Universidad Carlos III de Madrid

Escuela Politécnica Superior

GRADO EN INGENIERÍA DE SISTEMAS AUDIOVISUALES

APLICACIÓN DE TÉCNICAS DE APRENDIZAJE MÁQUINA PARA LA CARACTERIZACIÓN Y CLASIFICACIÓN DE PACIENTES CON TRASTORNO OBSESIVO COMPULSIVO

Autor: María López Bautista

Tutor: Vanessa Gómez Verdejo

Leganés, junio de 2013

Título

Aplicación de técnicas de aprendizaje máquina para la caracterización y clasificación de pacientes con Trastorno Obsesivo Compulsivo

Autor

María López Bautista

Tutor

Vanessa Gómez Verdejo

El TRIBUNAL

Presidente:

Secretario:

Vocal:

Realizado el acto de defensa del presente Trabajo Fin de Grado el día __ de _____ de 2013 en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acordando otorgarle la CALIFICACIÓN de:

Agradecimientos

La carrera ha sido como una maratón, con sus diferentes etapas: la euforia inicial, la fase de concentración en la que estás completamente centrado en la prueba, el momento en el que te ves incapaz de continuar y, al final, ese sprint final para la tan ansiada llegada a meta o, en nuestro caso, el trabajo fin de grado. Es entonces, cuando cruzas la meta y recibes tu recompensa, cuando echas la vista atrás y te acuerdas de todos los que te ayudaron a llegar hasta allí.

Gracias a Vanessa por aceptar el trabajo de orientar y supervisar el presente trabajo. Sin su paciencia y su esfuerzo no estaría aquí escribiendo estas palabras.

Gracias a mi familia, por convertirme en la persona que soy. Especialmente mis padres, Emilio y Elvira, por su apoyo incondicional en cada una de las decisiones tomadas y por el esfuerzo realizado para que yo tuviese las oportunidades que a ellos les fueron negadas.

Gracias a mis compañeros de la universidad, en particular a aquellos que siguen ahí, porque a su lado las decepciones fueron menores y las alegrías multiplicadas.

Y, por último, y no menos importante, gracias a mis amigos, de allí y de aquí, porque confiaron y creyeron en mí cuando yo había dejado de hacerlo.

A todos, muchas gracias.

Resumen

El siguiente Trabajo Fin de Grado se basa en el cada vez más habitual empleo de métodos de aprendizaje máquina con el fin de clasificar y caracterizar trastornos psiquiátricos. Concretamente, el sistema diseñado pretende acercarse al diagnóstico de TOC ('Trastorno Obsesivo Compulsivo') a través del análisis de imágenes de resonancia magnética (MRI).

El sistema diseñado tiene como objetivo plantear un algoritmo capaz de diagnosticar pacientes con TOC y, principalmente, capaz de caracterizar la enfermedad, detectando de manera automática las regiones neuroanatómicas relacionadas con el trastorno. Para ello, se empleará una arquitectura modular creada a partir de dos premisas fundamentales.

1. Análisis por áreas funcionales y/o neuroanatómicas. Cada imagen de resonancia magnética se divide en, aproximadamente, una centena de subconjuntos compuestos por vóxeles asociados a un área funcional o región neuroanatómica del cerebro. Así pues, el objetivo es aplicar un clasificador que facilite la selección de los conjuntos de vóxeles relevantes para la detección de la enfermedad.

2. Caracterización y fusión de áreas funcionales. El sistema utilizará métodos de selección de características sobre las salidas de los clasificadores el objetivo de obtener una selección automática de las áreas relevantes para el diagnóstico de la patología que estamos tratando. Asimismo, el último paso será el estudio de la relación que tienen las áreas entre sí mediante el uso de clasificadores, tanto lineales como no lineales.

Una vez desarrollado y aplicado el algoritmo, se aprovecharán los resultados tanto para comparar la clasificación de pacientes con los resultados previos obtenidos mediante métodos tradicionales [1], [2], como para analizar el patrón de áreas neuroanatómicas responsables del trastorno.

Abstract

This work is based on increasingly common use of machine learning methods in order to classify and characterize psychiatric disorders. Specifically, the designed system tries to be able to diagnose OCD (Obsessive-Compulsive Disorder) though the MRI (Magnetic Resonance Imaging) analysis.

The main system's goal is to construct an algorithm able to detect OCD patients and characterize the disease, detecting automatically neuroanatomical regions related to the disorder, supported on a modular arquitecture process with two fundamental principles.

1. Analysis of functional and/or neuroanatomical areas. Each MRI is divided into one hundred subsets composed of voxels associated to a functional area. Thus, the goal is to apply a classifier which facilitates the selection of the relevant voxels sets for the diagnosis of the disease.

2. Characterization and combination of functional areas. The system will use feature selection methods with the outputs of the first classifiers in order to get an automatic selection of the relevant areas for diagnosis of the pathology. The last step will use linear and no liner classifiers to analyze whether the different areas are interrelated.

Having the algorithm developed, we will use the results to compare the classifications of patients with previous results got by traditional methods [1], [2], and to analyze the pattern of neuroanatomical areas responsible for the disorder.

Índice general

Capítulo 1

Introducción

1.1. Descripción del problema

El presente Trabajo Fin de Grado se encuadra en uno de los campos en los que el procesado digital de imágenes está aportado una valiosa colaboración en los últimos tiempos: el análisis de imágenes médicas.

En la actualidad, debido a la gran cantidad de imágenes de las que dispone cada estudio, la extracción completa de información mediante el simple análisis visual de un experto médico resulta imposible; por lo que surge la necesidad de utilizar métodos de ayuda en ese proceso de obtención del conocimiento. Además, también debemos tener en cuenta que la experiencia y la subjetividad con que el experto revisa y analiza la imagen médica son factores influyentes en el diagnóstico final de una lesión. Por estas razones, la asistencia automática en la toma de decisiones en escenarios clínicos es un campo donde la penetración de las técnicas de aprendizaje máquina está cobrando bastante auge. Así, ya podemos encontrar aplicaciones de las mismas en la clasificación de células en los tejidos, la detección del cáncer de mama, la caracterización de la retinopatía, la detección de arritmias o fibrilaciones, etc.

En concreto, el trabajo queda acotado en un ámbito específico del análisis de imágenes médicas, limitándose al estudio de la detección de trastornos psiquiátricos, para lo cual nos centraremos en el análisis de imágenes de resonancia magnética, MRI ("Magnetic Resonance Imaging"). La elección del MRI resulta habitual para este tipo de investigaciones, dado que permiten analizar de manera no invasiva el cerebro humano.

De esta manera, en [3], [4] ya encontramos referencias del empleo de datos MRI funcionales para la detección de alzheimer, esquizofrenia y lesiones cerebrales traumáticas leves; mientras que en [5], [6] y [7] apuntan a la utilización de MRI estructural para la caracterización de depresión, autismo y alzheimer respectivamente. Sin embargo, a pesar de los avances ya existentes en este campo, en la mayoría de estas aproximaciones la solución resulta difícil de interpretar, condicionando su

aplicación en entornos reales. Por esta razón, en aplicaciones como la detección del trastorno obsesivo compulsivo (TOC), resulta de especial interés proporcionar una caracterización de la patología en térmicos anatómicos, aspecto al que hasta el momento no se ha prestado tanta atención, con el objetivo de mejorar la comprensión de la enfermedad y de facilitar al especialista realizar un diagnóstico adecuado y preciso.

Hasta el momento, el análisis de imágenes de resonancia magnética estructural de pacientes con TOC ha servido para descubrir algunas anomalías [1],[2], mientras que otros resultados preliminares [8] muestran la posibilidad de emplear información neuroanatómica para mejorar la detección.

Como guía de este trabajo se han utilizado líneas de investigación ya existentes del Departamento de Teoría de la Señal y Comunicaciones de la Universidad Carlos III de Madrid. Por una parte, sobre los métodos de clasificación máquina y su aplicación conjunta con técnicas de selección de características partimos de aproximaciones como las ya propuestas por algunos miembros del departamento en [9], [10], [11]. Y, por otra parte, respecto a la detección de patologías hemos trabajado en paralelo a líneas de investigación en desarrollo tales como técnicas para la detección de esquizofrenia [12] y/o para la detección de TOC [13], [14].

Así pues, en base a esos estudios previos, este trabajo propone alternativas, basadas en la aplicación de técnicas de procesado de datos, que contribuyan en la determinación de qué áreas del cerebro humano pueden resultar útiles para aproximarnos con éxito al diagnóstico de pacientes de TOC.

1.2. Objetivos

El propósito principal de este trabajo es conseguir de manera automática la clasificación de pacientes de TOC y la caracterización de la patología a partir de la caracterización de los pacientes mediante datos MRI estructurales.

El estudio se basa en la aplicación de herramientas de aprendizaje máquina, como son las Máquinas Vectores Soporte, SVM (“Support Vector Machine”), [15],[16], con las que, partiendo de ejemplos etiquetados, un algoritmo aprende a discernir si una muestra, no empleada en dicho aprendizaje, pertenece a una clase determinada. El empleo de estas técnicas facilitará la clasificación de pacientes, pero acorde a nuestros objetivos se prestará especial atención a su uso en combinación con métodos de selección de características [17],[18], los cuales nos permitirán caracterizar el

trastorno y localizar las áreas neuroanatómicas responsables del mismo, posibilitando analizar su correspondencia con los resultados previos adquiridos a través de métodos tradicionales de diagnóstico.
De este modo, en este trabajo nos centraremos en conseguir una mejor caracterización del trastorno.

Para implementar nuestro sistema, hemos de tener en cuenta algunos inconvenientes que surgen cuando se trabaja con imágenes MRI. Algunos de esos factores escapan de nuestro alcance, como el coste y dificultad de adquirir este tipo de imágenes, mientras que otros como la alta dimensionalidad de los datos y el problema de las diferencias cerebrales a nivel estructural entre diferentes individuos pueden ser mitigados en gran parte a través de técnicas que de preprocesado o métodos que permitan definir qué información del cerebro es relevante para detectar el trastorno. Así pues, con el objetivo de simplificar el procesado de los datos, vamos a suponer que la información en el cerebro se encuentra dispersa. Esta hipótesis implica que únicamente determinados conjuntos de vóxeles (áreas funcionales) del cerebro contendrán información relevante para la clasificación de los pacientes, e identificarlas nos servirá para obtener resultados más concretos y precisos entorno la detección de la patología.

Partiendo de esa hipótesis inicial, el sistema se basará en la idea principal de construir un clasificador capaz de diferenciar entre los sujetos control y los pacientes que padecen trastorno obsesivo compulsivo. Para ello el primer paso será la aplicación de un clasificación SVM a cada área que permita indicar si en ella existen o no indicios de la patología. Con esa información, por medio de la fusión de esos resultados, la aplicación de las técnicas de selección de características nos permitirá esbozar un diagnóstico global (presencia o no de la enfermedad) basado en las áreas más relevantes en la toma de decisión. Manteniendo como objetivo principal caracterizar la patología, es decir, identificar las regiones cerebrales relacionadas con ella, las áreas relevantes serán aquellas que formen el subconjunto que minimice la tasa de error del clasificador final.

De este modo, el trabajo comienza por una implementación básica del sistema completo y va incorporando y evaluando el impacto en las prestaciones del sistema de alguna de las siguientes contribuciones:

- Diseño tanto de los clasificadores empleados sobre cada área como el utilizado para una fusión final del sistema. El objetivo es valorar las posibilidades que ofrecen tanto SVMs lineales como no lineales en ambas posiciones de la cadena evaluando su rendimiento.

- Empleo de diferentes métodos de selección de características. Se analizarán las prestaciones obtenidas para la caracterización y el análisis de la relevancia de las diversas áreas funcionales. En esta dirección, se aplicará el método de eliminación recursiva de variables [19], técnicas de selección automática de características mediante regularizaciones de dispersión o técnicas que permitan prefijar el número de áreas que se desean emplear en el modelo final [20], [11].

En definitiva, nuestras metas son diseñar diferentes experimentos entorno al algoritmo base para validar el modelo y analizar los resultados con el propósito de extraer conclusiones dentro del ámbito del aprendizaje máquina empleado en el procesado de imágenes de resonancia magnética.

1.3. Estructura del texto

Como el propio título indica, en este apartado se van a esbozar breves descripciones del contenido de cada capítulo en el que se dividen los conceptos teóricos, la estructura y la validación experimental de las diferentes soluciones propuestas para la detección y caracterización del TOC.

Los datos utilizados en el presente trabajo han sido extraídos de imágenes de resonancia magnética previamente pre-procesadas para su análisis estadístico. El Capítulo 2 se centrará en introducir tanto la naturaleza de estos datos y el procedimiento necesario para su obtención como el tratamiento que se les aplica antes de ser utilizados como datos de entrada en nuestros experimentos. De esta forma, esta sección nos permitirá comprender de qué información disponemos y cómo podemos utilizarla en nuestro beneficio.

Como el título de la memoria indica, el objetivo es usar el aprendizaje máquina para clasificar y categorizar pacientes de TOC. Por ello, hemos considerado necesario una exposición teórica de las técnicas que, combinándolas, conformarán nuestras propuestas, acercándonos a ese objetivo. Así, mientras que el Capítulo 3 trata de describir qué es el aprendizaje máquina y qué distintas posibilidades nos ofrece para poder utilizarlas en los diferentes puntos del sistema según la naturaleza del problema; el Capítulo 4 recorre, de manera general, los tipos de selección de características existentes y, dentro de cada uno, qué métodos hemos desarrollado para su inclusión en el proyecto.

Una vez completados los conceptos teóricos, en el Capítulo 5 se pasa a desarrollar las propuestas llevadas a cabo y los experimentos que las han secundado.

Así pues, esa sección está dividida en tres bloques. El primero es una introducción en la que se expondrán los aspectos que tienen en común los experimentos, el segundo supone una descripción de los experimentos y sus validaciones y una exposición de los resultados obtenidos a través de ellos, y el tercero contiene la valoración de las prestaciones de cada uno de los algoritmos propuestos. Desde el punto de vista ingenieril, esto se llevará a cabo comparando los valores de error proporcionados por cada uno de los sistemas.

El Capítulo 6 recopila las conclusiones finales obtenidas tras estudiar el rendimiento de cada uno de los sistemas propuestos. Asimismo, en este mismo capítulo también se encuentran las líneas de investigación futuras que pueden seguirse para extender el trabajo realizado.

Para terminar, en el último de los capítulos se halla la información sobre la planificación que ha conllevado este trabajo y el presupuesto total que supone teniendo en cuenta tanto costes personales como materiales.

Capítulo 2

Datos de MRI

2.1. Introducción

El trastorno obsesivo compulsivo, TOC u OCD ("Obsessive-compulsive disorder"), es un trastorno de ansiedad caracterizado por pensamientos intrusivos, recurrentes y persistentes que producen inquietud, aprensión, temor o preocupación, y conductas repetitivas. Tiene consecuencias significativas en la actividad natural del enfermo a nivel familiar, profesional y social. Se trata de un trastorno psiquiátrico crónico con una presencia del 2% en la población mundial [14].

Actualmente, manuales como el DSM-IV (Diagnostic and Statistical Manual of Mental Disorders) establecen las pautas a seguir para diagnosticar la enfermedad. El diagnóstico mediante psicoterapia y el tratamiento farmacológico se perfilan como opciones habituales para evitar recaídas de los pacientes. Sin embargo, para conocer en profundidad el trastorno, se han llevado a cabo investigaciones apoyadas en técnicas de neuroimagen estructural (MRI, TAC) y funcional (fMRI, PET) cuyos resultados parecen apuntar a causas genéticas y alteraciones cerebrales como origen de la aparición de la enfermedad.

Así pues, en la búsqueda de una aproximación al diagnóstico de dicho desorden psiquiátrico vamos a partir de modelos neurobiológicos sólidos ya existentes basados en el análisis de resonancias magnéticas estructurales (sMRI) etiquetadas en función de si el individuo presenta o no el trastorno. Particularmente, las imágenes a utilizar han sido tratadas reutilizando los principios del VBM ("Voxel Based Morphometry"), puesto que, probablemente, es el método más utilizado para caracterizar anormalidades estructurales del cerebro en pacientes con TOC y otros trastornos psiquiátricos como alzheimer, esquizofrenia, etc. VBM es una técnica común y automatizada que permite el estudio de diferencias focales en la anatomía cerebral para establecer correlaciones entre el volumen cerebral y variables clínicas, usando una aproximación estadística paramétrica

2.2. Resonancia Magnética Estructural

La Resonancia Magnética (RM) es una técnica relativamente nueva, utilizada desde principios de la década de los 80, en la obtención de imágenes útiles en el diagnóstico de enfermedades. Se basa en el uso de ondas magnéticas y de radio, por lo que no hay exposición a rayos X u otras formas de radiación perjudicial. En ese punto reside su mayor ventaja, en ser una técnica de neuroimagen que permite analizar imágenes en vivo del cerebro humano de manera no invasiva.

La RM es eficiente si queremos obtener información sobre la estructura y composición del cuerpo a analizar. Se apoya en un fenómeno físico basado en utilizar las propiedades que adquieren los núcleos atómicos al bañarlos en ondas de radio frecuencia tras haberlos sometidos a un intenso campo magnético. Generalmente, la información que ofrecen gira entorno a las estructuras del propio cuerpo, datos que también pueden ser extraídos a través de radiografías, ecografías o TAC. Asimismo, también permite estudiar problemas que ante esas técnicas son transparentes.

Los dispositivos que debe contener el aparato de RM para poder llevar a cabo el proceso son los siguientes: un imán fijo y potente, imanes secundarios, bobinas emisoras y receptoras de ondas de radio, una antena para recibir señales emitidas por los tejidos, y un ordenador capaz de representar imágenes o analizar el espectro, cuyo fruto será la base de nuestro estudio.

El resultado que nos facilita el software del mecanismo de RM se trata de una imagen 3D formada por vóxeles. Al igual que sucede con un píxel en un espacio de dos dimensiones, el vóxel es la mínima unidad de volumen que constituye un objeto 3D. Un vóxel, por tanto, es un elemento de volumen que contiene información gráfica asociada a un punto en un espacio tridimensional.

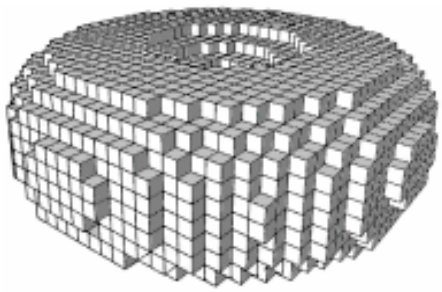

Figura 1. Representación tridimensional construida por unidades elementales de volumen o vóxeles

Así pues, cada una de esas unidades de volumen estará relacionada con un dato numérico generado por el software del aparato de RM basándose en los parámetros

de entrada ajustables de la máquina. De manera genérica, la información que nos proporciona estará vinculada con la densidad de la sustancia que ocupa ese punto en el espacio real. De esta manera, podremos distinguir diferentes tipos de sustancias según el valor de su densidad. A modo de ejemplo, la siguiente imagen nos permite diferenciar entre materia gris ("GM", Grey Matter), materia blanca ("WM", White Matter) y líquido cefalorraquídeo ("CSF", Cerebrospinal fluid).

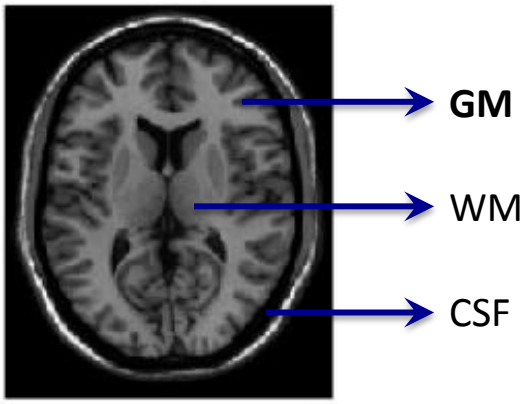

Figura 2. Ejemplo de imagen de resonancia magnética cerebral segmentada en tres tipos de sustancias: materia gris ("GM", Grey Matter), materia blanca ("WM", White Matter) y líquido cefalorraquídeo ("CSF", Cerebrospinal fluid)

2.3. Pre-procesado de datos

La adquisición de las imágenes cerebrales que conforman la base de nuestros experimentos se dio lugar a partir de dos selecciones. La primera de 86 pacientes de TOC pertenecientes al servicio ambulatorio del Departamento de Psiquiatría, Hospital Universitario de Bellvitge, Barcelona; que se seleccionaron tras diagnosticar el trastorno por la presencia de síntomas típicos y su persistencia. Para el caso de los sujetos control, se conformó un grupo de 86 personas que superaron la entrevista 'Structured Clinical Interview for DWV-IV (SCID)' para descartar cualquier desorden psíquico.

En relación a las imágenes resulta significativo tener en cuenta que cuando se realizan análisis de grupos se asume que todos los cerebros individuales están analizados de tal manera que cada vóxel se localiza en la misma región anatómica para todos los sujetos. Tras esta suposición y anterior al análisis estadístico que conlleva la VBM, los datos son sometidos a una serie de varias etapas de pre-procesamiento para simplificar el análisis estadístico posterior. Dicho pre-procesado previo conlleva la segmentación de las imágenes en tejidos, la normalización de las mismas a un espacio anatómico común y el filtrado de las imágenes.

A continuación detallaremos las etapas que conforman el tratamiento previo de las imágenes al que han sido sometidos los datos de resonancia magnética que empleamos en nuestra propuesta [21].

- **Análisis de ROIs**

  En primer lugar es necesario dividir el cerebro en áreas o regiones de interés, ROI (Regions Of Interest), para, posteriormente, poder analizar por separado su relevancia en el diagnóstico.

  Para poder dividir el cerebro en áreas es necesario utilizar un criterio, que puede variar según qué objetivo tengamos. Un ejemplo de división puede ser en base a las áreas de Brodmann [22], neurólogo alemán que dividió la corteza cerebral en 52 regiones según los tejidos que poseen células nerviosas.

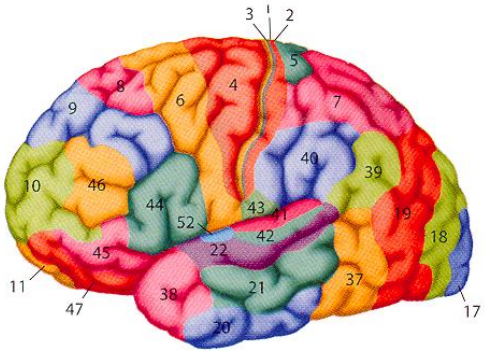

Figura 3. Áreas de Brodmann

  En nuestro caso, la división de áreas de la que partimos ha sido extraída de la fragmentación anatómica del cerebro llevada a cabo por el MNI ("Montreal Neurological Institute") en uno de sus estudios [23].

- **Segmentación**

  La segmentación de imágenes tiene como principal objetivo transformar su representación en otra más útil y fácil de analizar.

  Este proceso servirá para dividir una imagen de resonancia en las diferentes materias que componen el cerebro humano. Para ello, se estima la probabilidad que tiene cada vóxel de ser de un tipo de materia u otro mediante técnicas del procesado de señal como son las mezclas de gaussianas.

  A partir de la información de la resonancia se estima la función de probabilidad que siguen los datos con una mezcla de gaussianas, consiguiendo una distribución como la de la figura. En ésta cada gaussiana se asocia a un tipo de materia y cada vóxel adquiere un peso mayor o menor dentro de cada gaussiana, obteniendo así la probabilidad de pertenecer a un tipo u otro de materia.

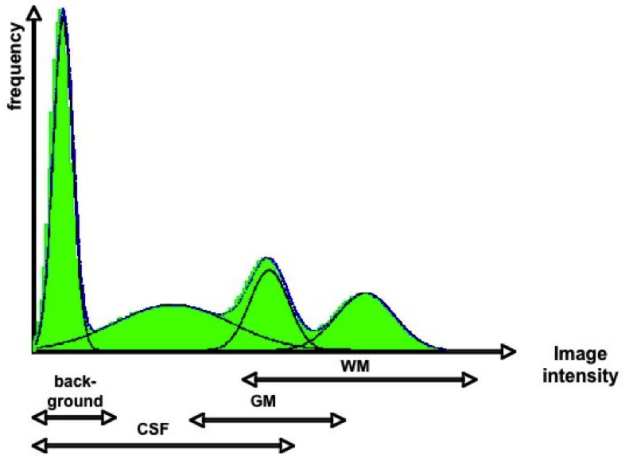

Figura 4. Función de transformación RM-mapas de probabilidad de presencia de una sustancia

Particularmente, en los datos que vamos a usar cada vóxel se caracteriza por su probabilidad de contener materia gris.

- **Normalización**

  El tercer paso, y el más importante para que los datos de entrada del experimento pertenezcan a un mismo espacio y puedan ser comparados, es llevar a cabo una normalización de todos los datos.

  Cada cerebro tiene diferentes forma y características, pero existen regularidades compartidas por todos los cerebros sanos. Así pues, la normalización intenta ajustar la anatomía de los sujetos a un espacio normalizado definido por una plantilla del cerebro; como por ejemplo, el 'MNI brain' (Montreal Neurological Institute), sin incurrir en distorsiones significativas. Para conseguirlo es necesario aplicar diversas transformaciones (no lineales) que adapten cada imagen al marco común.

  Entre sus ventajas cabe destacar que permite informar e interpretar localizaciones espaciales de manera consistente, y que los resultados pueden ser generalizados para una gran población, pudiendo ser comparados a través de estudios y sujetos. Los inconvenientes son que reduce la resolución espacial y puede introducir errores derivados de la interpolación.

- **Suavizado espacial**

  El suavizado es la última de las transformaciones que se aplica a las imágenes antes de su empleo en el sistema. El proceso se basa en aplicar un filtro gaussiano a los datos de muestreo antes de la reconstrucción. Esto implica que muchos de los datos adquiridos son descartados como producto del suavizado, facilitando el análisis estadístico.

  En concreto, los datos utilizados en este proyecto han sido tratados con un suavizado de 4mm de anchura.

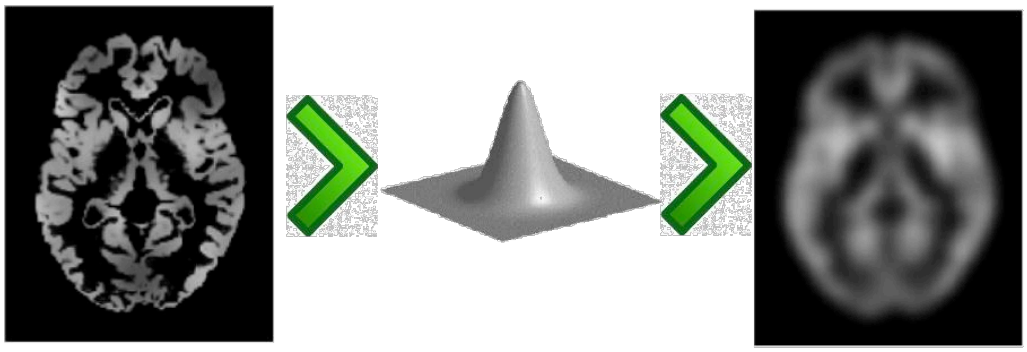

Figura 5. Efecto del suavizado sobre una imagen de RM

Capítulo 3

Aprendizaje Máquina

3.1. Introducción

Podemos considerar 'aprender' como una forma determinada de hacer que un elemento, hardware o software, que interactúa con su entorno sea capaz de auto-configurarse para solucionar, con el tiempo, una tarea que se le encarga repetidamente. En la actualidad, la tecnología disponible ya ha permitido conformar sistemas que resuelven diversos tipos de tareas con una limitación: los problemas ya estaban previstos. Así pues, para superar dicha barrera necesitamos dotar a esos sistemas de inteligencia, la cual residirá en su adaptabilidad, en que sea capaz de observar su entorno y aprender de él. El objetivo es que la máquina pueda integrar nuevos conocimientos.

El Aprendizaje Máquina o Aprendizaje Automático, ML ("Machine Learning"), es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar métodos que permitan a cualquier dispositivo electrónico aprender y cambiar su comportamiento de manera autónoma basándose en la experiencia [24]. Resulta común que esta disciplina acabe solapándose con la estadística, puesto que ambas se sustentan en el análisis de datos, centrándose el Aprendizaje Máquina en la complejidad computacional del problema.

El Aprendizaje Máquina ofrece técnicas muy efectivas para el descubrimiento de patrones en grandes volúmenes de datos, jugando un rol fundamental en áreas tales como bioinformática, previsión meteorológica, inteligencia de negocios o reconocimientos faciales.

Dentro de esta disciplina podemos encontrar diferentes tipos de aprendizajes según su taxonomía. El que nos concierne es el denominado aprendizaje supervisado. Esta tipología se basa en algoritmos cuyo fin es establecer una correspondencia entre las entradas y las salidas del sistema. Un ejemplo de ello es el problema de clasificación, que es aquel que nos compete, pues el sistema trata de clasificar una

serie de vectores utilizando una entre varias categorías. La base del conocimiento de un algoritmo supervisado son datos etiquetados anteriormente.

Nuestros experimentos van a girar entorno a un sistema de aprendizaje máquina denominado Máquinas de Vectores Soporte. En concreto, utilizaremos un subconjunto de algoritmos de aprendizaje supervisado que, a partir de datos etiquetados, sean capaces de aprender de manera automática para llevar a cabo clasificaciones binarias.

3.2. Máquinas Vectores Soporte

La Máquina de Vectores Soporte, SVM ("Support Vector Machine") tiene su origen en 1964, cuando Vapnik y Lerneren la desarrollaron en los laboratorios AT&T. Pero no fue hasta los años 90 cuando, a través de diversas publicaciones (Boser, 1992; Cortes y Vapnik, 1995) se convirtió en un éxito por su mejor generalización respecto a las Redes Neuronales, proporcionando buenas prestaciones con pocos datos y un espacio de entrada de dimensión alta.

Las SVM fueron originalmente ideadas para la resolución de problemas de clasificación binaria en los que las clases eran linealmente separables. Su objetivo final es construir un hiperplano que separe de forma óptima los datos de las dos clases que comparten espacio. La selección óptima del hiperplano pasa por elegir el que mejor represente el límite entre los dos conjuntos, es decir, el que maximiza el margen entre las muestras de cada clase respecto a la frontera de separación [25]. Así pues, la intención es que a partir de M observaciones podamos encontrar una función tal que:

$$\mathrm{f}: \mathbf{x}_i \longrightarrow \{\pm 1\} \qquad (3.1)$$

$$\text{donde } \mathbf{x}_\mathrm{i} \in \Re^\mathrm{N} \; \forall \; \mathrm{i} = 1, \ldots, \mathrm{N}$$

En la búsqueda del hiperplano se realiza un entrenamiento de los datos con el fin de que la máquina generalice bien para que los datos nuevos (que no han participado en el entrenamiento) sean clasificados correctamente. Para lograrlo se busca seleccionar unos parámetros de la máquina que proporcionen un equilibrio entre minimizar el coste empírico sobre el conjunto de entrenamiento y mantener una correcta generalización de la máquina [26].

A continuación se presentarán los principios básicos de las diferentes técnicas SVM. En primer lugar, diferenciaremos en dos casos de clasificación lineal según si los

datos de los que partimos se pueden separar o no linealmente. En segundo lugar se introducirá la versión no lineal de la máquina.

3.2.1. Caso linealmente separable

Se considera que dos conjuntos son linealmente separables cuando existe un hiperplano capaz de clasificarlos con error cero. El objetivo, por tanto, es encontrar la frontera de decisión de máximo margen, entendiendo margen como la distancia entre ella y los datos, que nos permite no cometer ningún error en la clasificación.

En la figura que se muestra a continuación se puede observar un caso binario linealmente separable donde el hiperplano maximiza el margen (margen 2).

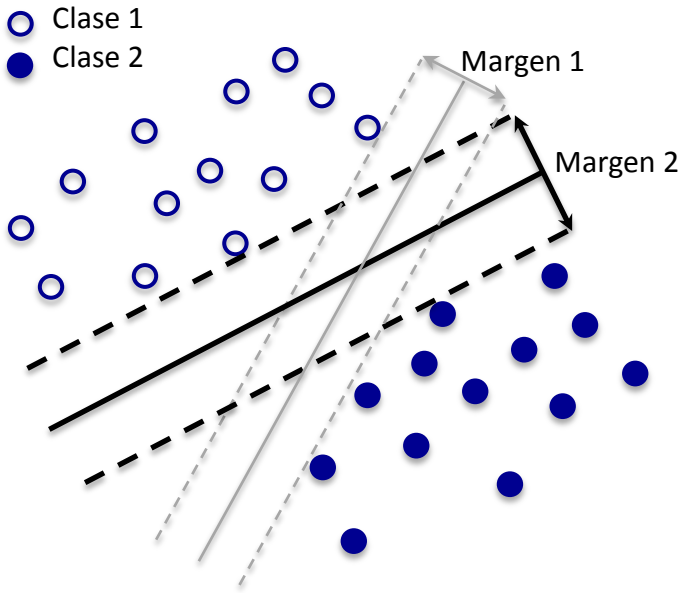

Figura 6. SVM en el caso separable

Matemáticamente partiremos de un conjunto de N muestras definido como $\{\boldsymbol{x}_i, y_i\}$, i $= 1, \ldots,$ N y etiquetado según $y_i \in \{-1,1\}, y_i \in \Re^D$, donde D es la dimensión de los vectores que contienen los datos.

Un clasificador es lineal si su función de decisión puede expresarse mediante una función lineal en x. Así pues, podremos representar el hiperplano que separa los puntos como:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + \mathrm{b} = 0 \tag{3.2}$$

donde x son los datos, w el vector normal al hiperplano y b representa el sesgo, constante que indica la posición del plano respecto al origen de coordenadas.

En definitiva, la clasificación consiste en determinar en qué zona del hiperplano está el punto a clasificar. Por ello, y basándonos en la ecuación (3.2), podemos redefinir el clasificador con las expresiones (3.3) y (3.4), que son resumidas en (3.5).

$$\mathbf{x}_i \cdot \mathbf{w} + b > +1 \text{ para } y_i = +1 \tag{3.3}$$

$$\mathbf{x}_i \cdot \mathbf{w} + b < -1 \text{ para } y_i = -1 \tag{3.4}$$

$$y_i\{\mathbf{x}_i \cdot \mathbf{w} + b\} \geq 1 \tag{3.5}$$

En el caso que nos ocupa, problema linealmente separable, habrá infinitos hiperplanos que cumplan esa definición, por lo que es necesario incluir una nueva condición que nos permita encontrar aquel con mayor margen, consiguiendo así maximizar la distancia entre datos y frontera de decisión. Para conseguirlo tendremos en cuenta los puntos para los que se cumple la igualdad, obteniendo dos hiperplanos paralelos entre sí y paralelos a su vez a la frontera de decisión.

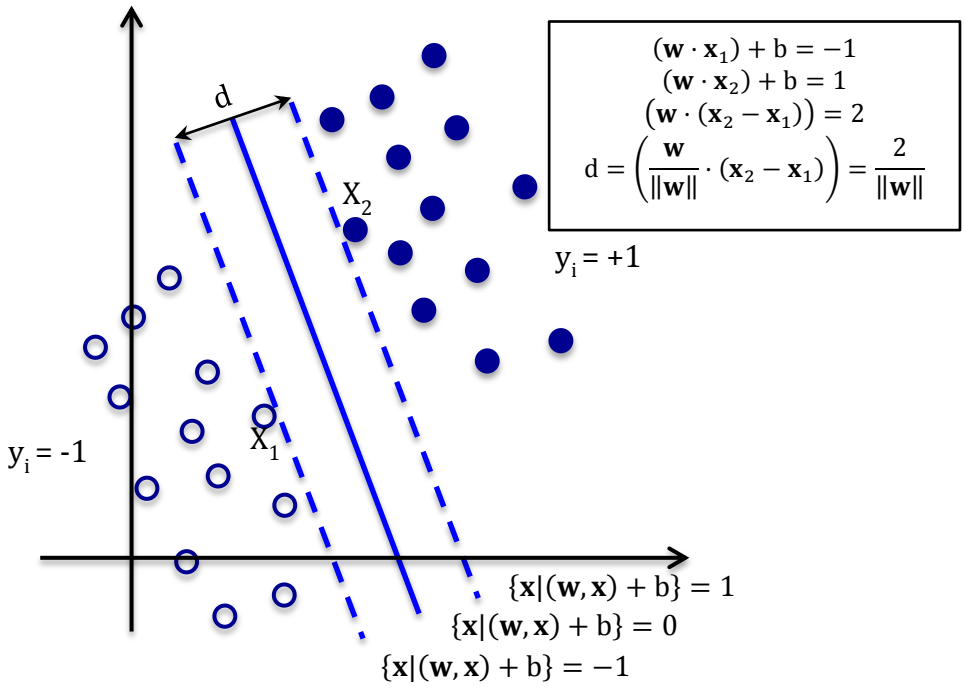

Figura 7. Hiperplano de separación óptimo normalizado para conjuntos linealmente separables

Teniendo en cuenta la anterior figura, la distancia de cada hiperplano al origen es $\frac{1}{\|\mathbf{w}\|}$, de lo que podemos deducir que el margen del hiperplano es $\frac{2}{\|\mathbf{w}\|}$. De este modo, para maximizarlo será suficiente con minimizar la norma de **w** sujeto a la restricción (3.5):

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.6}$$

$$\text{s.t. } y_i\{\mathbf{x}_i \cdot \mathbf{w} + b\} \geq 1$$

Llegados a este punto, el empleo de los multiplicadores de Lagrange permite representar el hiperplano como combinación lineal de las propias muestras. Incorporándolos (3.3) y (3.4) la expresión a minimizar queda en (3.7), la cual, tras ser derivada e igualada a 0, nos permite deducir las condiciones (3.8) y (3.9):

$$\frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i (y_i(\mathbf{w}^T\mathbf{x_i} + b) - 1) \tag{3.7}$$

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \tag{3.8}$$

$$\sum_i \alpha_i y_i = 0 \tag{3.9}$$

Aplicando estas restricciones podemos definir la función de clasificación:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i (\mathbf{x_i}\mathbf{x}) + b\right) \tag{3.10}$$

En este contexto se denominan Vectores Soporte a aquellos datos que se encuentran justo a la distancia del margen y son los únicos datos necesarios para definir la frontera de decisión debido a que sus multiplicadores de Lagrange, a diferencia del resto, serán distintos de 0.

3.2.2. Caso linealmente no separable

El caso anterior sólo funcionaría cuando una frontera lineal puede delimitar los dos conjuntos de datos. Cuando no es posible por el solapamiento de las clases la solución pasa por encontrar el hiperplano que cometa el menor número de errores posibles.

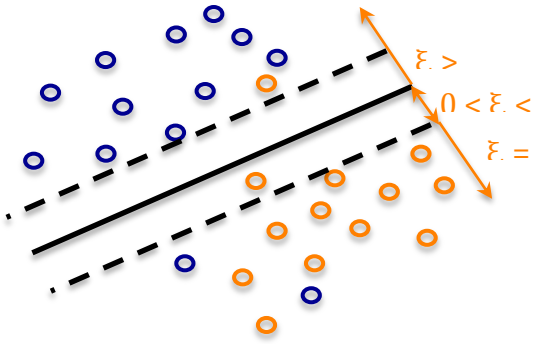

Figura 8. SVM en el caso no separable

En la búsqueda de dicho hiperplano, la restricción (3.5) se completa incluyendo unas variables de relajación con el fin de controlar el error y penalizar las muestras mal clasificadas, quedando:

$$y_i\{\mathbf{x}_i \cdot \mathbf{w} + b\} \geq 1 - \xi_i \tag{3.11}$$

con

$$\xi_i > 0\ , \forall i \tag{3.12}$$

Con estas condiciones, se cumplirá que en las muestras bien clasificadas $0 < \xi_i < 1$, dependiendo cómo de cerca estén de la frontera, y en las mal clasificadas $\xi_i > 1$.

Así pues, la ecuación a minimizar se modificará quedando de la siguiente manera:

$$\min_{\mathbf{w},\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{3.13}$$

$$\text{s.t. } y_i\{\mathbf{x}_i \cdot \mathbf{w} + b\} \geq 1, \qquad \xi_i > 0 \ \ \forall i$$

donde C es una constante que regula el compromiso entre la maximización de la distancia y la minimización de errores.

Del mismo modo que en el caso linealmente separable, utilizando multiplicadores de Lagrange, la ecuación (3.13) se convierte en (3.14). Derivándola respecto a cada una de sus variables obtendremos las restricciones del caso anterior ((3.8),(3.9)) y (3.15), lo que implica que los multiplicadores quedan limitados a $0 \leq \alpha_i \leq C$.

$$L_{w,b,\propto,\mu} = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\alpha_i(y_i\ (\mathbf{w}^T\mathbf{x_i} + b) - 1 + \xi_i) \tag{3.14}$$

$$C - \mu_i - \alpha_i = 0 \tag{3.15}$$

3.2.3. SVM no lineal

En ocasiones, debido a la distribución de los datos sobre el espacio de entrada, la solución lineal desarrollada en los apartados anteriores no proporciona un buen resultado. En esos casos necesitaremos trazar fronteras de clasificación no lineales, para lo cual recurrimos a proyectar los datos a un espacio de mayor dimensión donde la solución lineal sea válida, operación facilitada por la SVM no lineal.

La SVM no lineal puede ser interpretada como una generalización del hiperplano de decisión. Se basa en aplicar una transformación al espacio de entrada, $\phi(\mathbf{x})$, con el fin de obtener un espacio de características donde las muestras sí sean separables. Así pues, la formulación es básicamente la misma, puesto que solo habrá que realizar el producto escalar en el espacio de características en lugar de en el espacio de entrada ($\mathbf{x} \longrightarrow \phi(\mathbf{x})$).

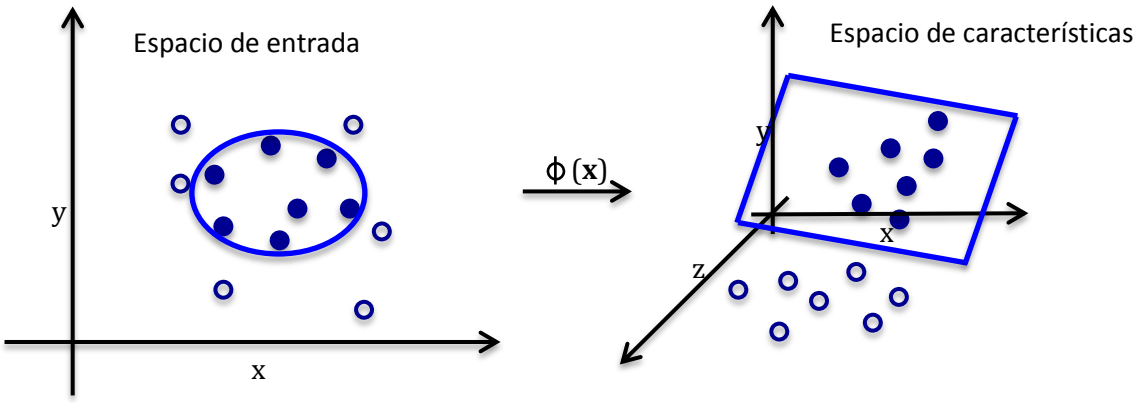

Figura 9. Transformación espacial del espacio de entrada

A partir de esa idea, es importante tener en cuenta que para construir una SVM en el espacio resultante éste debe ser un Espacio de Hilbert, donde la proyección del espacio de entrada se lleva a cabo mediante una función núcleo (kernel) que se puede definir como:

$$k(\mathbf{x}, \mathbf{x}_i) =< \phi(\mathbf{x}), \phi(\mathbf{x}_i) > \tag{3.16}$$

La cual, a su vez, facilita aplicar el algoritmo de las SVM, puesto que permite medir distancias entre patrones sobre el espacio de características sin conocer la transformación ϕ. Así pues, cualquier función $k(\mathbf{x},\mathbf{x}_i)$ válida, que lo será si cumple el Teorema de Mercer, nos servirá para aplicar el procedimiento.

Tras la transformación y las anteriores consideraciones, finalmente la función de decisión se redefine como:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i \cdot k(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{3.17}$$

Existen diversas funciones kernel que permiten adaptar la SVM a los diferentes conjuntos de muestras con el fin de obtener mejores resultados según la naturaleza de los mismo. Entre las más comunes podemos destacar:

- *Lineal*

$$k(\mathbf{x}, \mathbf{x}_i) = \mathbf{x} \cdot \mathbf{x}_i \tag{3.18}$$

- *Polinómica*

$$\mathrm{k}(\mathbf{x}, \mathbf{x}_\mathrm{i}) = (\gamma \mathbf{x} \cdot \mathbf{x}_\mathrm{i} + \mathrm{C})^{\alpha} \tag{3.19}$$

- *Gaussiana*

$$\mathrm{k}(\mathbf{x}, \mathbf{x}_\mathrm{i}) = \exp(-\gamma |\mathbf{x} - \mathbf{x}_\mathrm{i}|^2) \tag{3.20}$$

donde α es el rango del polinomio, C un coeficiente y γ es una contante de proporcionalidad.

Un factor a considerar en esta técnica es el coste computacional, el cual se verá afectado por el incremento en el número de dimensiones debido a la transformación. Aún así, generalmente, su impacto se ve suavizado por el hecho de que no es necesario trabajar en el espacio de características, sino que basta con conocer su producto escalar (3.16).

Capítulo 4

Métodos de selección de características

4.1. Introducción

Una de las principales dificultades de utilizar métodos de aprendizaje máquina sobre datos MRI es que cada volumen de datos recogido contiene decenas de miles de vóxeles. Es decir, la dimensión de cada conjunto de información es muy alta cuando se compara con el número de datos del experimento, cuyo orden de magnitud oscila entre las decenas y los cientos de imágenes. Esa gran diferencia en la dimensionalidad de datos y número de observaciones disponibles afectará al rendimiento de la máquina en términos de generalización. Además, se conoce que la información relevante para el análisis que queremos realizar está contenida únicamente en unos pocos vóxeles, por lo que poder detectarlos nos permitirá ganar en interpretabilidad y nos ayudará a analizar como una patología como el TOC se distribuye en el cerebro.

Por estas razones, los métodos de selección de características se convierten en una herramienta clave para el procesado de la información de la que disponemos. Los objetivos de los mismos serán:

- Reducir la dimensionalidad de los datos perdiendo la menor cantidad de información posible. Deben encontrar el equilibrio entre un número de datos con el que resulte sencillo trabajar y el hecho de que la pérdida de información no sea relevante, es decir, no conlleve pérdida de esos vóxeles importantes para la detección que buscamos.

- La selección de voxels resultante proporcione información neuroanatómica y clínica de interés para el estudio del TOC.

- Su aplicación suponga una carga computacional asequible (menos capacidad de almacenamiento y procesos de entrenamiento rápidos).

- Y, si es posible, permita mejorar las prestaciones del proceso de aprendizaje.

4.2. Métodos de selección de características

Cuando se nos presenta una tarea de aprendizaje a resolver y partimos de un conjunto de datos de entrada en el que el número de variables a tener en cuenta es muy elevado, el empleo de métodos de selección de características debe facilitar la búsqueda de un subconjunto cuya información sea más relevante en nuestra tarea.

Siguiendo la clasificación proporcionada por [27], los métodos de selección de características se pueden agrupar en tres categorías:

1. Filtros ("Filters"): Usan medidas de relevancia para analizar lo predictiva que es cada variable o subconjunto de ellas. Debido a que la tarea suele ser fija, la selección será independiente de ella y vendrá dada en función de los subconjuntos de características existentes. Generalmente combinan los criterios de relevancia con algoritmos de búsqueda o, directamente, generan un ranking de variables según esas medidas. Son los algoritmos que menor carga computacional suponen.

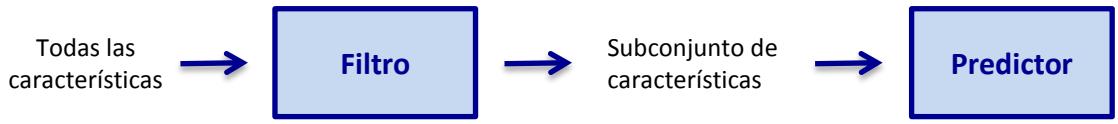

Figura 10. Esquema del algoritmo tipo 'filter'

2. Envolventes ("Wrappers"): Emplean el aprendizaje máquina para seleccionar subconjuntos de características. Su proceso pasa por entrenar el modelo con diferentes subconjuntos de características y valorar, a partir de la tasa de error de un clasificador, cada uno de ellos. Trabajar junto a los predictores les permite retroalimentarse de las variables de salida de ellos. Normalmente proporcionan la mejor selección para el modelo particular con el que se está trabajando, pero su carga computacional es muy alta.

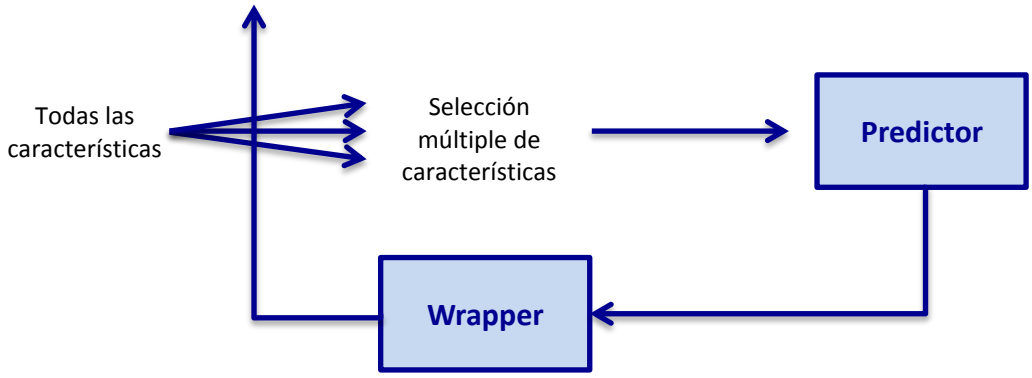

Figura 11. Esquema del algoritmo tipo 'embedded'

3. Incrustados ("Embedded"): Algoritmos integrados en la predicción. Se diferencian de los otros métodos en su interacción con el aprendizaje máquina, jugando la estructura de éste un papel fundamental en la selección. Realizan el proceso de selección de características de forma paralela al entrenamiento de la máquina, seleccionando aquellas que consideran de mayor utilidad para la clasificación. Al estar vinculados al aprendizaje, serán específicos para cada máquina.
   En términos de complejidad computacional se encuentran entre los filtros y los métodos envolventes. Y son menos propensos al sobreajuste que los métodos tipo 'wrapper'.

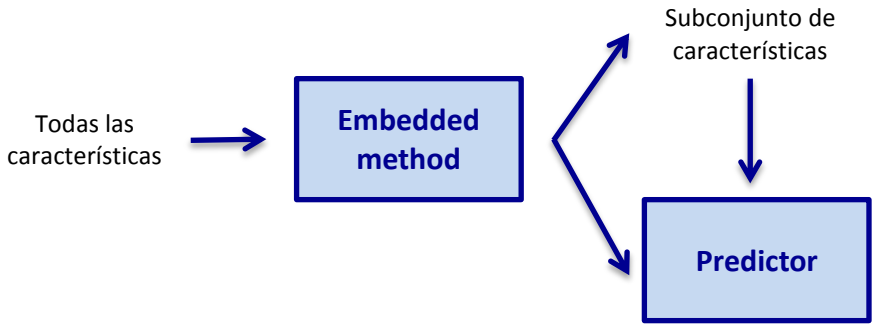

Figura 12. Esquema del algoritmo tipo 'embedded'

En los experimentos descritos en el siguiente capítulo se van a utilizar los métodos de selección de características que a continuación planteamos y encuadramos en cada una de las clases expuestas.

4.2.1. Ranking de variables

Dentro de la categoría de filtros uno de los métodos más extendidos es la realización de clasificaciones o rankings de importancia de las características. Estos

rankings son realizados teniendo en cuenta únicamente la relación existente entre la característica que está siendo analizada y la variable supervisada.

Para llevar a cabo el ranking de las características según sus relevancias individuales se deben llevar a cabo dos sencillos pasos: obtener esas relevancias individuales y ordenar las características según su relevancia.

De esta manera, el punto más complicado del algoritmo reside en la generación de un coeficiente o peso para cada uno de nuestros atributos o características. Para ello se suelen aplicar distintos heurísticos derivados de medidas de divergencia entre funciones de distribución. En función de la métrica utilizada, los coeficientes de mayor valor serán asignados a los atributos más relevantes del problema o a los de menor trascendencia.

Así pues, el objetivo es que en las primeras posiciones de los rankings de atributos tengamos características que despejan gran parte de la incertidumbre del problema que queremos resolver, mientras que en las zonas finales estarán aquellas que parecen no tener apenas relación con el problema y cuyo uso en el aprendizaje máquina pueden resultar más perjudicial que beneficioso.

En definitiva, este método consiste en obtener un ranking que nos permita ir utilizando de 1 a N características de mayor a menor relevancia para encontrar la cantidad con la que se optimiza el problema de aprendizaje máquina posterior.

Un aspecto a tener en cuenta de este tipo de algoritmos es que al realizarse las medidas de manera univariante, es decir, valorando únicamente la característica de manera individual, no se tienen en cuenta las posibles relaciones o redundancias entre las características. De esa forma, existe información que no manejan, si las variables individualmente irrelevantes pueden resultar relevantes al combinarse con otras y si las variables individualmente relevantes pueden presentar redundancia al unirse a otras.

Una de las mayores ventajas de este tipo de medidas es la rapidez y eficiencia con la que se realiza la clasificación, sobre todo cuando el número de dimensiones de los datos es mucho mayor al número de datos disponible. Sin embargo, si la búsqueda resulta muy detallada puede dar lugar a sobreajuste.

4.2.2. Wrappers

Otro de los algoritmos que se utilizan en los experimentos es un ‘wrapper’ natural. Esa categoría de métodos de selección de características consiste en combinar

la búsqueda de características en el espacio de atributos con el algoritmo de aprendizaje, evaluando diferentes conjuntos de atributos y escogiendo el más adecuado. Se basan en la validación cruzada.

Facilitan hallar el conjunto de características más útil, pero tienden a sobreajustar y su coste computacional está por encima del método anterior, pues repite el mismo proceso una y otra vez hasta dar con el mejor conjunto de atributos.

Igual que anteriormente, en el ranking de variables, la decisión a tomar para desarrollar el algoritmo era cómo generar las relevancias individuales de las características; en este caso, lo que se debe decidir previamente es la manera en que se van a seleccionar los subconjuntos de características con los que evaluar el entrenamiento de la máquina.

Consiguientemente, lo que para definir nuestro 'wrapper' debemos elegir es cómo vamos a formar los subconjuntos. En nuestro caso, los datos de cada área compondrán una característica y lo que evaluaremos serán subconjuntos formados por todas las características excepto por una. Es decir, cada vez que entrenemos nuestra máquina lo haremos con los datos de todas las áreas menos una, evaluando así el sistema con N subconjuntos donde N es el número de áreas funcionales en las que dividimos nuestros datos de MRI.

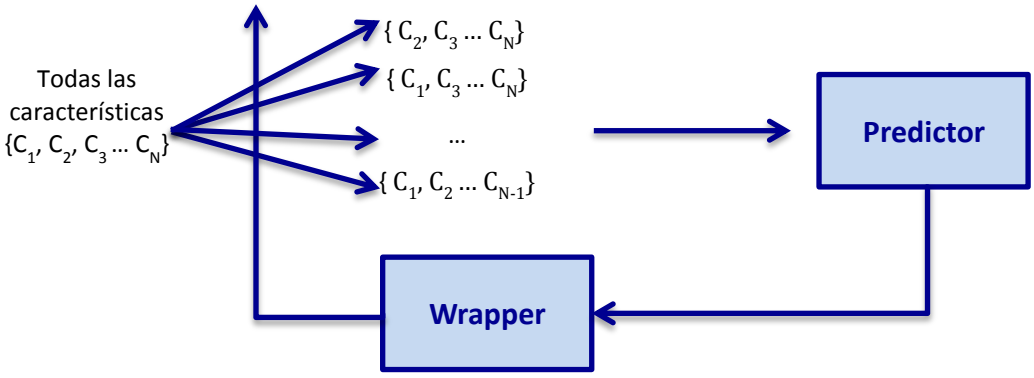

Figura 13. Esquema del algoritmo tipo 'wrapper' con los subconjuntos de características seleccionados

De esa forma, una vez completados todos los ciclos, consideraremos que la característica menos relevante y, por tanto, prescindible, será aquella cuya extracción haya generado la menor tasa de error. Así pues, excluyendo ese atributo quedaría seleccionado el subconjunto de N-1 áreas óptimo.

Una vez seleccionado el subconjunto óptimo se repetirá el proceso para ir eliminando variable a variable hasta quedarnos con una única región.

4.2.3. Recursive Feature Elimination

Una vez desarrollados los algoritmos de tipo 'filter' y de tipo 'wrapper' que se van a utilizar en el proyecto, la 'Eliminación recursiva de características' es el primero de los dos métodos desarrollados que se encuadran dentro del tipo 'embedded', puesto que mide la utilidad de la característica a la hora de incluirla entre el subconjunto óptimo para resolver el problema.

El RFE es un método patentado por Isabelle Guyon en 2002, [19], que nace como una manera de encontrar relaciones discriminatorias en conjuntos de datos clínicos. Este método permite tanto filtrar las características irrelevantes sobre las relacionados con la enfermedad, como identificar patrones respecto a las mismas.

RFE es un método de selección de características desarrollado entorno a las SVM que trata de encontrar el mejor subconjunto de tamaño $\sigma < n$ (n = dimensiones de entrada) a través de una eliminación hacia atrás '*gradual*' [18]. Así pues, se basa en el principio básico de la eliminación hacia atrás ('Backward Elimination'), el cual consiste en, partiendo de todo el conjunto de características, eliminar un atributo en cada iteración hasta llegar a quedarse con la característica más trascendental en la clasificación que se persigue.

En definitiva, el RFE trata de seleccionar, a partir de la salida del entrenamiento de una SVM, el conjunto de características que den lugar al mayor margen de separación de clases en el problema binario que nos ocupa. Ese objetivo se logra extrayendo en cada iteración aquella característica cuya ausencia a la hora de entrenar una SVM provoca la mayor disminución del margen.

Así pues, por cada iteración el sistema funcionará de la siguiente manera:

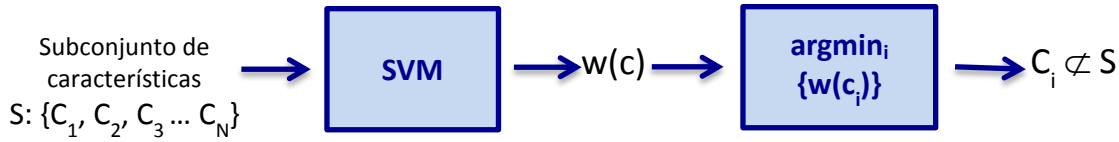

Figura 14. Esquema de cada iteración del RFE

Al final, lo que este método permite obtener, si se llevan a cabo tantas iteraciones como número de características tenemos, es una lista ordenada de subconjuntos de atributos.

4.2.4. SVM norma 1

En la resolución de problemas de clasificación es habitual emplear términos de regularización sobre los parámetros del modelo si se desea obtener una solución dispersa. Así pues, el segundo de los métodos de tipo 'embedded' que se utilizan en el proyecto se trata del modelo de SVM norma 1, en el cual la selección de atributos puede ser vista como un problema de optimización.

La formulación de este método surge de sustituir el algoritmo SVM estándar, es decir, con norma 2, por un algoritmo que minimice la norma 1 del vector de pesos **w**. Para ello, reformularemos el problema de la siguiente manera:

$$\min_{\mathbf{b},\mathbf{w},\xi} \ \|\mathbf{w}\|_1 + C\sum_i \xi_i \qquad (4.1)$$
$$\text{s.t. } y_i\{\mathbf{x}_i \cdot \mathbf{w} + b\} \geq 1 - \xi_i, \qquad \xi_i > 0 \ \forall i$$

La regularización respecto a la norma 1 lo que permite es evitar que los coeficientes **w** tomen valores altos (reduciendo así la varianza) y, por consiguiente, evitar las soluciones sobreajustadas.

Por otra parte, el modelo SVM norma 1 posee dos propiedades que son las que posibilitan la selección automática de variables, produciéndose lo que se llama una solución dispersa (pocos coeficientes no nulos), especialmente en problemas de alta dimensionalidad. Esas propiedades son las siguientes:

1. Al presentar una discontinuidad en el origen, la penalización de coeficientes cercanos a 0 es mucho mayor que para el caso de norma 2, pues realiza un truncado automático de los coeficientes a 0. Como consecuencia de ello, la norma 1 tiene la propiedad de seleccionar variables automáticamente, todas aquellas cuyo peso sea distinto de 0; mientras que la norma 2 proporcionaría valores pequeños de esos pesos, pero nunca llegarían a ser 0.

2. Regulando el valor del parámetro de regularización se puede controlar el número de coeficientes no nulos.

A continuación se muestra una representación del problema de estimación cuando se emplean ambas regularizaciones: norma 1 y norma 2; donde los contornos entorno al origen muestran la curva de nivel proporcionada por un parámetro de regularización de valor dado, mientras las elipses representan las curvas de nivel de la función de error de mínimos cuadrados.

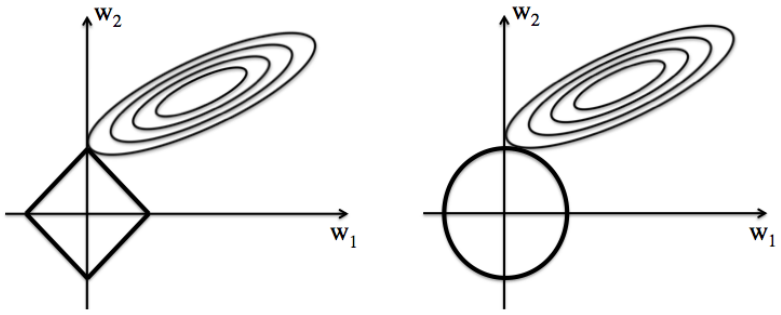

Figura 15. Comparativa curvas de penalización para norma 1 (izquierda) y norma 2 (derecha)

A pesar de las diversas ventajas que este modelo ofrece, en ciertos casos también tiene limitaciones. Uno de estos inconvenientes ocurre cuando las variables de entrada están altamente correladas en el conjunto de datos y todas ellas son relevantes para la variable de salida. En este caso la norma 1 tiende a seleccionar solo algunos de esos atributos y a reducir el resto a 0. Otra limitación surgiría cuando la dimensión de datos de entrada es mayor al número de datos (M > N), puesto que la norma 1 seleccionaría como máximo N atributos.

Capítulo 5

Experimentos y resultados

5.1. Descripción de los experimentos

En este primer apartado, previo a la descripción de los diferentes sistemas propuestos, se muestran las técnicas empleadas para el entrenamiento y evaluación de los experimentos, es decir, cómo se dividen los datos y cómo se utilizan en las diferentes etapas de cada proceso para conseguir dos condiciones necesarias en cualquier sistema:

1. La validación de los parámetros del sistema de clasificación.

2. Una evaluación del sistema que nos permita valorar cómo de positivo está siendo su aplicación en el asunto que nos ocupa: el diagnóstico de los pacientes de TOC y la caracterización de la patología.

Es decir, este apartado sirve para describir la estructura general que se va a usar en los experimentos respecto al empleo óptimo de los datos, sin detallar qué técnicas de aprendizaje máquina y de selección de características usan cada uno.

5.1.1. Conjunto de datos de entrada

Como se describe en el Capítulo 2 de este trabajo, los datos a partir de los cuales se desarrollan los diferentes experimentos se corresponden con un conjunto de imágenes MRI ya procesadas cuya información se encuentra descompuesta en las siguientes estructuras:

- **Matriz MRI_DATA**: Matriz de dimensiones (num_voxels)x172 en la que cada columna representa al cerebro de un sujeto. Los valores de esta matriz se corresponden a la probabilidad de que el vóxel al que representa sea materia gris.

num_voxels es el número de datos que tenemos de cada cerebro en 3D pero redimensionado en una columna. En nuestro caso, 482.315 dimensiones.

- **Matriz AREAS:** Matriz de dimensiones 116x2 que nos facilita la división por áreas de los datos de la matriz MRI_DATA. Por un lado, el número de filas se debe a que disponemos de 116 regiones cerebrales en las que fragmentar los datos de las MRI. Por otro lado, el número de columnas se corresponde a los dos datos que tenemos de cada área. La primera columna contiene los códigos que van a representar a cada una de las áreas, mientras que la segunda encierra los nombres que se le asignan a cada una de las regiones de la posición homónima de la primera columna.

- **Vector AREAS_VOXELS**: Vector de longitud (num_voxels) cuyos datos relacionan cada vóxel con la región cerebral a la que pertenece. Los valores de este vector podrán ser 0 u oscilar entre los diferentes valores que tomaba la primera columna de la matriz AREAS. Cualquier valor no nulo representará el área a la que pertenece; mientras que los valores nulos simbolizan que ese vóxel no se incluye en ninguna de las áreas que vamos a evaluar.

- **Vector Y**: Vector de longitud 172 que contiene las etiquetas para los cerebros disponibles. El valor de dichas etiquetas indican si la imagen pertenece a un individuo del grupo de sujetos de control o, en cambio, si corresponde a un paciente. Concretamente, se ha escogido el valor +1 para los 86 pacientes de TOC y el -1 para designárselo a los sujetos de control.
  Los valores de Y no son datos de entrada como tal, que estemos cargando a partir de un fichero; sin embargo, resultan primordiales para el proceso que lleva a cabo cualquiera de los experimentos.

Una vez definidos los datos de los que partimos, el siguiente paso es conocer como se van a organizar para llevar a cabo el entrenamiento, la validación y la evaluación de cada modelo.

5.1.2. Entrenamiento, validación y test: Validación cruzada

Para evaluar cualquier sistema es necesario el cálculo de un error de test sobre un conjunto de datos no empleado durante el entrenamiento. Sin embargo, debido al escaso número de datos del que se dispone, no resulta factible reducir aún más el conjunto de entrenamiento. En estas ocasiones, en la que la disponibilidad de cantidad de datos es reducida, la solución para el problema pasa por utilizar la validación cruzada.

La validación cruzada es una técnica que nos permite garantizar que los resultados de nuestro análisis estadístico son independientes de la división entre entrenamiento y test, es decir, permite evitar el sobreajuste. Consiste en repetir el proceso sobre diferentes divisiones de conjuntos de entrenamiento y test y calcular la media aritmética de la media de evaluación. Resulta útil en entornos donde se quiere estimar cómo de preciso es un modelo predictivo.

En nuestro caso, se han utilizado dos métodos de validación cruzada: Leave One Out (LOO) y Double Leave One Out (Double LOO). Resultan procesos computacionalmente costosos, pero capaces de reducir al mínimo el número de datos no disponibles para el entrenamiento, esencial en nuestras circunstancias. El funcionamiento de ambos se expone a continuación.

5.1.2.1. Leave One Out (LOO)

El LOO consiste en dividir los N datos disponibles en dos subconjuntos. El primero de ellos, de entrenamiento, estará formado por todos los datos menos uno, es decir, (N-1) datos; mientras que el subconjunto de test constará del dato restante. De esta manera, del LOO se llevan a cabo N iteraciones, dejando en cada una de ellas un elemento diferentes para el subconjunto de test, que evalúa el modelo creado por los restantes datos de entrenamiento.

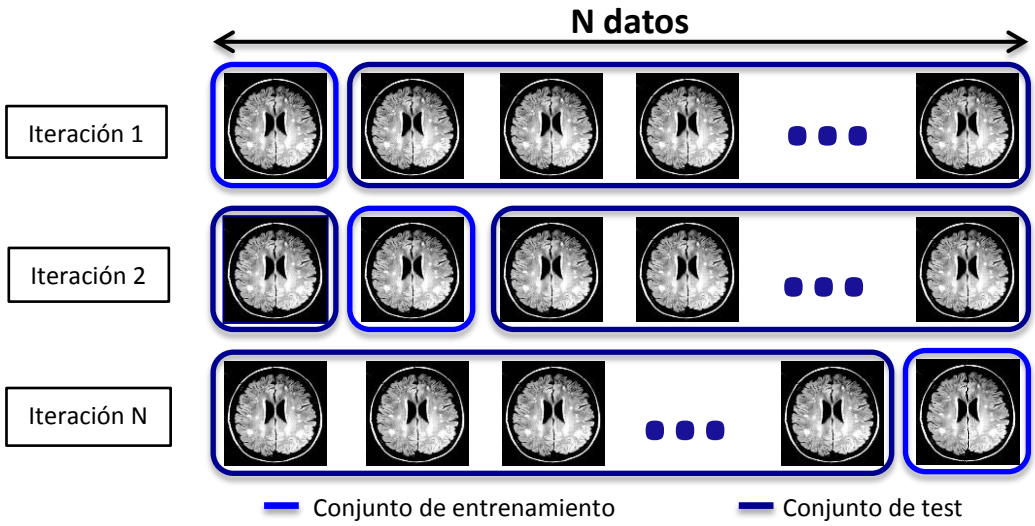

Figura 16. Conjuntos de entrenamiento y de test en cada iteración del LOO

A través de las etiquetas disponibles para cada dato es posible deducir si ha habido error o no en la decisión, y la media aritmética de esos errores equivale a la tasa de error de test del sistema a evaluar.

Datos X: x_i para i = {1,2…N} Etiquetas Y: y_i para i = {1,2…N}

Para cada x_i

- $S_{entrenamiento}$ = X excepto x_i
- $S_{test} = x_i$
- Proceso de entrenamiento -> decisión $d(x_i)$
- error(i) = Comparar y_i con $d(x_i)$

tasa_error_sistema = media_aritmetica(error)

Figura 17. Esquema de la evaluación del sistema a través del LOO

5.1.2.2. Double LOO

El doble LOO sigue el mismo principio que el proceso anterior, con la diferencia de que no sólo sirve para independizar la tasa de error del conjunto de entrenamiento, sino que nos permite llevar a cabo la validación de parámetros.

En primer lugar realiza la misma división de datos que el LOO simple, obteniendo un subconjunto de entrenamiento y un subconjunto de test. La diferencia aparece justo en el siguiente paso, antes de entrenar la SVM con el subconjunto correspondiente, pues para ello es necesario validar los parámetros del sistema. Así pues, previa al proceso de entrenamiento, existe otra división de los datos. Esta segunda división nos proporciona un proceso anidado en el que un LOO va internamente dentro de otro, separando el subconjunto de entrenamiento en dos: entrenamiento y validación. El primero estará compuesto por (N-2) elementos, mientras que el de validación quedará constituido por un único dato. Así, con cada valor del parámetro a validar se creará un bucle en el que en cada iteración el dato de validación sea diferente, dando lugar al final a una tasa de error cuyo valor mínimo servirá para elegir el valor óptimo para el parámetro.

Datos X: x_i para i = {1,2…N} Etiquetas Y: y_i para i = {1,2…N}

Para cada x_i

- $S_{entrenamiento}$ = X excepto x_i=s_j para j = {1,2…N-1}
- $S_{test} = x_i$

  Para cada valor de parámetro a validar

  Para cada s_j

  - $S_{entrenamiento2}$ = $S_{entrenamiento}$ excepto s_j
  - $S_{val} = s_j$
  - Proceso de entrenamiento -> decisión $d(s_j)$
  - error_val = Comparar y_j con $d(s_j)$

  tasa_error_val = media_aritmetica(error_val) -> min(tasa_error_val) -> param
- Proceso de entrenamiento con $S_{entrenamiento}$ y param -> decisión $d(x_i)$
- error(i) = Comparar y_i con $d(x_i)$

tasa_error_sistema = media_aritmetica(error)

Figura 18. Esquema de la evaluación del sistema a través del LOO

En definitiva, el doble LOO, como su propio nombre indica consiste en realizar un LOO sencillo sobre el subconjunto de entrenamiento derivado de la división en el primero de los LOO.

5.1.3. Validación de parámetros del sistema

La construcción de cualquier sistema de clasificación requiere definir previamente una serie de parámetros desconocidos de los cuales dependerá el resultado. Para conseguir con el valor con el cual el sistema presenta sus prestaciones óptimas, el procedimiento a seguir es validar el problema barriendo diferentes rangos para los parámetros en libertad.

A lo largo de las técnicas de aprendizaje y de los métodos de selección de características descritos anteriormente se encuentran los diversos parámetros que ha sido necesario validar y que se recogen a continuación:

- *C*: Valor utilizado en cada una de las SVM. Este parámetro es una constante que regulariza el compromiso entre el número de errores de entrenamiento permitidos y la búsqueda de la maximización del margen.

- Γ: gamma se refiere al parámetro utilizado para construir el kernel gaussiano en cada una de las SVM no lineales.

- C_{norma1}: Parámetro utilizado en el modelo SVM norma 1. Su función es la determinación de los pesos, por tanto, su valor permitirá controlar el número de coeficientes nulos.

Al final, para el funcionamiento del sistema se emplean los parámetros óptimos, es decir, se utilizan aquellos que en la búsqueda de parámetros han dado lugar a una menor tasa de error de validación.

5.2. Métodos globales

A lo largo del desarrollo de los diferentes experimentos a exponer a continuación, éstos han sido organizados en dos categorías bien diferenciadas: métodos globales y métodos secuenciales.

Los que, primeramente, nos ocupan son los métodos globales. Esta denominación se aplica a todos aquellos experimentos que tienen una arquitectura única, es decir, aquellos en los que se consigue un diagnóstico y una caracterización de la enfermedad a través de un tratamiento global de los datos, sin incurrir en una separación de los procesos en etapas o en procedimientos específicos para cada fracción de los datos.

En estos métodos resulta de vital importancia reducir al máximo su sensibilidad a la dimensionalidad de los datos de entrada puesto que, debido a que se van a emplear el total de los voxels, ésta tenderá a ser alta, implicando un coste computacional elevado y peligro de sobreajuste.

5.2.1. SVM (lineal y no lineal)

El primer sistema propuesto para la detección de TOC se trata del procedimiento más genérico y global posible. Consiste en entrenar un único clasificador SVM empleando todos los voxels contenidos en las imágenes cerebrales de MRI. Se trata de un método sencillo basado únicamente en el aprendizaje máquina, y en el cual todavía no se introduce ninguno de los métodos de selección de características descritos en el capítulo anterior.

Se utilizan los métodos LOO y doble LOO para llevar a cabo el entrenamiento, la validación y la evaluación del sistema. Así pues, cada iteración, considerando un LOO donde el parámetro de libertad ya se encuentra validado, tendría la siguiente forma.

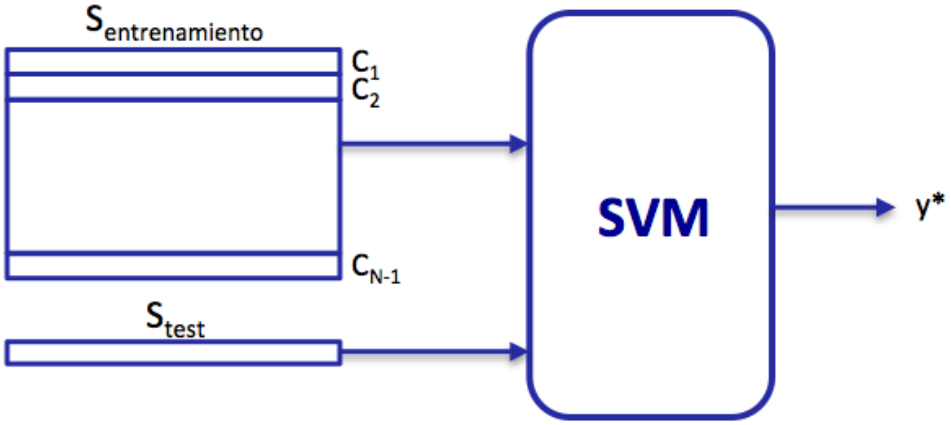

Figura 19. Esquema de cada iteración del LOO en el sistema de clasificación sin dividir los datos de entrada en regiones cerebrales

Para completar el sistema se debe seleccionar qué tipo de función de kernel va a utilizar la máquina, es decir, elegir el tipo de SVM. Como ya se ha explicado en el Capítulo 3, esa elección está determinada por la relación existente entre los datos que

se utilicen de entrenamiento. Como es lógico, se empleará un clasificador SVM lineal en el caso de que la relación entre los voxels del cerebro sea lineal; mientras que si no es así, será necesario utilizador un clasificador SVM no lineal en el que definiendo una función kernel se consiga obtener un espacio lineal en un espacio transformado.

En nuestro caso, dado que a priori se desconoce la relación existente entre los voxels del cerebro, emplearemos ambos clasificadores, pudiendo deducir al final cuál es el tipo de SVM idóneo para este problema, pues será aquel que presente el menor error de clasificación.

Los resultados de este esquema nos servirán como solución inicial a nuestro problema, pudiendo utilizarlos como referencia en el posterior análisis de todos los experimentos. Este punto de partida nos facilitará concluir si los sistemas de clasificación de mayor complejidad, donde se incluyen técnicas más avanzadas, suponen mejores prestaciones o, por lo contrario, el coste computacional que implican resulta innecesario.

5.2.1.1. SVM lineal

El primero de los experimentos a analizar se trata de aplicar una SVM lineal a todo el grueso de nuestros datos. Para ello, se utilizará un doble LOO en el que el primer bucle sirva para extraer la tasa de error del sistema y el segundo para validar el parámetro C que conlleva la utilización de la SVM lineal.

- **Independencia de C**

Como se indicaba en el apartado de la validación de parámetros, C es una constante que nos permite regular el compromiso entre los errores y la maximización de la distancia entre datos e hiperplano. Para tener conciencia de cómo varía nuestro error según dicho parámetro, el experimento ha utilizado el un barrido logarítmico amplio que nos permita valorar su incidencia en los resultados. El rango utilizado ha sido el siguiente: un barrido que comprenda 10 valores distanciados logarítmicamente entre 10^{-3} y 10^{3}.

En la siguiente gráfica se muestran los resultados, en cuanto a error de validación se refiere, resultado de utilizar cada uno de nuestros valores del barrido de *C*. El eje de abscisas simboliza los 10 valores que C tenía asignados, mientras que en el eje de ordenadas se representa el error de validación. Para interpretar el dibujo correctamente es necesario tener en cuenta que estamos representando el valor medio del error de validación en función de C,

considerando como margen en los el que puede variar la desviación estándar del mismo.

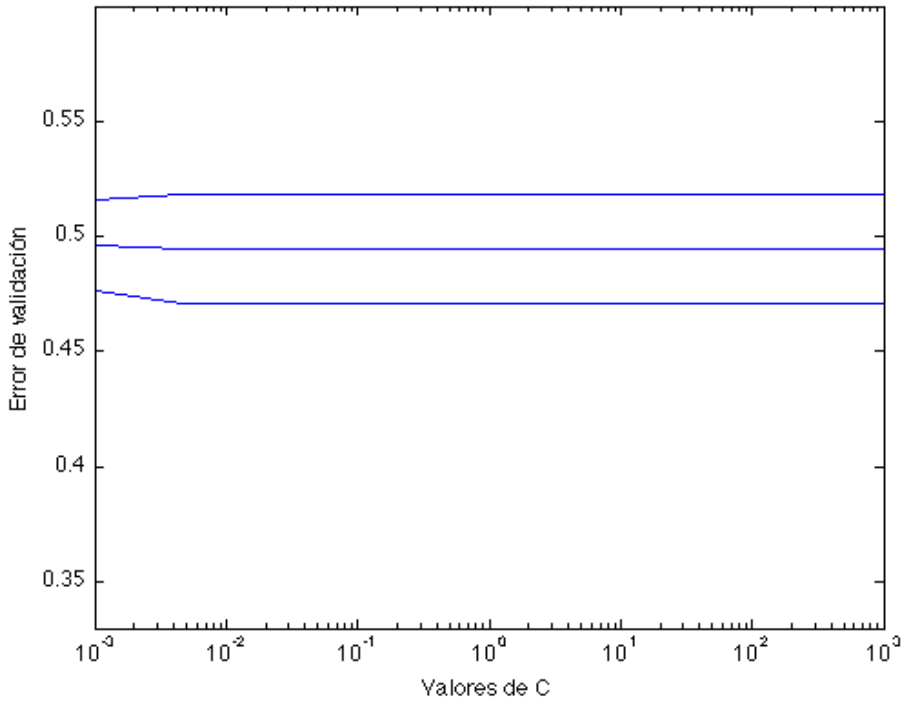

Figura 20. Error de validación del experimento SVM lineal en función del barrido logarítmico a C

A partir del segundo valor del barrido el error no varía sea cual sea el C utilizado. Esta observación nos permite concluir que será útil en éste y todos los experimentos restantes: debido a las características de los datos que estamos utilizando, para la SVM lineal es innecesario validar el parámetro C, pues su variación no altera en ningún caso los resultados significativamente. Por tanto, desde este punto, siempre que utilicemos una SVM lineal utilizaremos un C fijo de 10^{-2}. La selección de un valor de los más bajos en nuestro barrido se debe a que computacionalmente es menos costoso trabajar con valores pequeños.

- **Resultados**

Finalmente, ejecutando el LOO sencillo con $C=10^{-2}$ para poder extraer la tasa de error del sistema, el resultado que nos proporciona este primer experimento es el siguiente error de clasificación: 41.28%.

5.2.1.2. SVM no lineal

El segundo de los experimentos a examinar se trata de replicar el anterior pero esta vez utilizando una SVM no lineal. Esta particularidad se diferencia de la

anterior en que la máquina utilizará un kernel gaussiano en lugar de una función lineal y que esta vez, además de validar *C,* deberemos también validar el otro parámetro que conlleva la utilización de la SVM no lineal: Γ, empleado para construir el kernel gaussiano. Así pues, será necesario de nuevo la utilización del doble LOO.

- **Validación de parámetros: *C* y Γ**

Tal y como ocurría en el caso anterior, primero vamos a analizar la validación de parámetros de la SVM no lineal para comprobar si, como ya ha ocurrido, alguno de los dos parámetros a validar puede ser fijado, simplificando la algorítmica del problema y reduciendo el coste computacional del sistema, pues al fijar un parámetro de libertad lo que conseguimos es extraer un bucle LOO del sistema. Así pues, para valorar uno por uno la influencia que los parámetros tienen en nuestro sistema fijaremos el otro y barreremos un rango logarítmico con aquel que estamos evaluando.

En primer lugar nos disponemos a validar el parámetro C. Para ello fijamos Γ y barremos el mismo rango que en el caso anterior: 10 valores con distribución logarítmica entre 10^{-3} y 10^{3}. A continuación, la Figura 21 representa de nuevo los errores de validación derivados de barrer el rango de C, lo que nos sirve para comprobar la influencia del parámetro en el sistema.

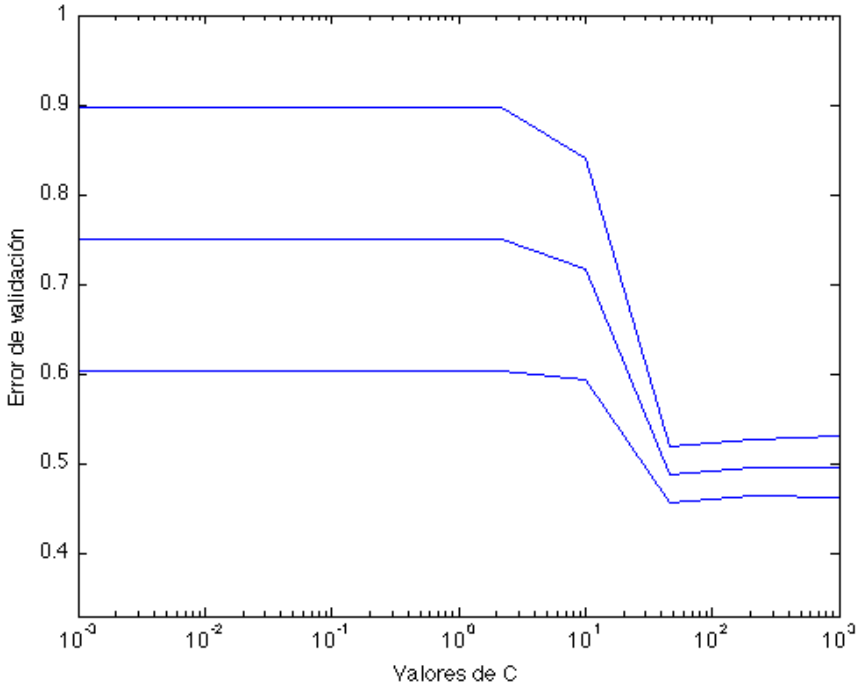

Figura 21. Error de validación del experimento SVM no lineal en función del barrido logarítmico a C con valores entre 10^{-3} y 10^{3}

La gráfica nos permite deducir de manera directa que es a partir del octavo valor asignado a C cuando éste es influyente en el resultado, reduciendo drásticamente los errores de salida. Por ello, antes de llegar a una conclusión final, lo que hemos hecho es utilizar un rango de C de valores más elevados para analizar su influencia. Así pues, este segundo rango es un barrido logarítmico de 10 valores entre 10^2 y 10^5.

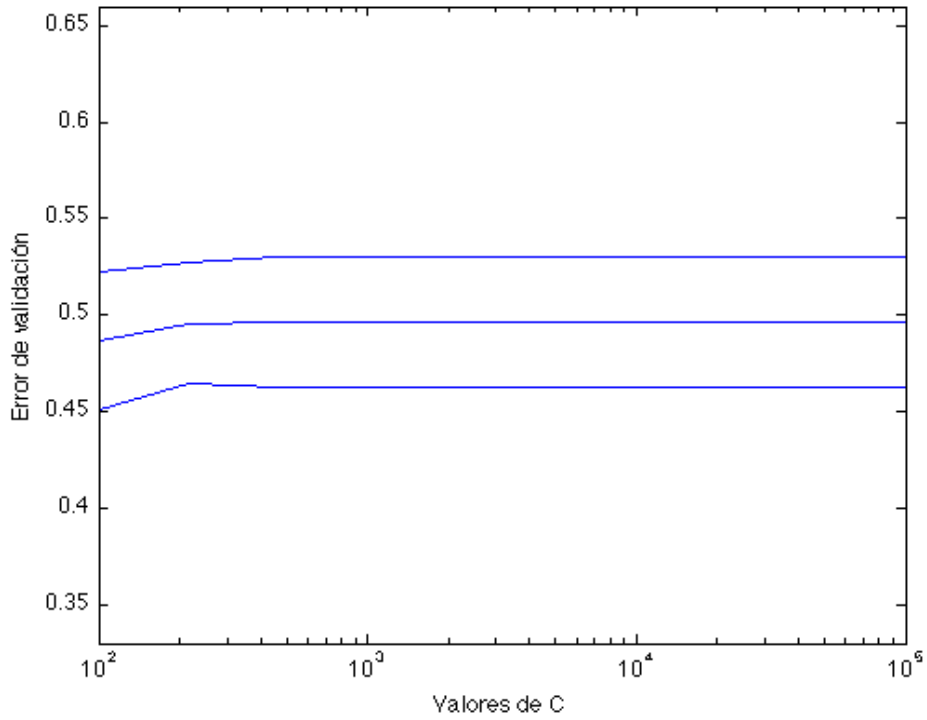

Figura 22. Error de validación del experimento SVM no lineal en función del barrido logarítmico a C con valores entre 10^2 y 10^5

A través de esta gráfica podemos deducir que, tal y como ocurría con la SVM lineal, C no influye en el resultado mientras que su valor sea lo suficientemente elevado para estabilizar la salida. Aún así, la siguiente tabla recoge el porcentaje de veces que cada valor asignado a C ha sido seleccionado como óptimo.

Valor de C	Óptimo en % casos
100	70.3%
215.4	16.9%
464.2	12.8%
Resto	0%

Tabla 1. Porcentaje de casos en los que cada valor asignado al parámetro de libertad C es seleccionado como óptimo

Esta circunstancia nos permitirá prescindir de su validación fijando el valor del parámetro en 100. Además, nos permite concluir que, dado que la independencia entre el resultado y el valor de C es debido a las características de los datos con los que estamos trabajando, para cualquier clasificador SVM no lineal que vayamos a utilizar en experimentos posteriores la validación de C será innecesaria.

Por otra parte, el procedimiento a seguir para validar Γ es el mismo. Una vez fijado C a 10^2 completamos un barrido logarítmico para Γ que nos permita evaluar su influencia en la tasa de error del sistema global. El rango elegido para ello va a depender directamente de los datos de entrada, por lo que utilizaremos el barrido logarítmico que tome 25 valores entre 10^{-3} y 10^3 y construimos el rango de gamma como el producto de esas constantes por el valor mediano de los datos de entrada.

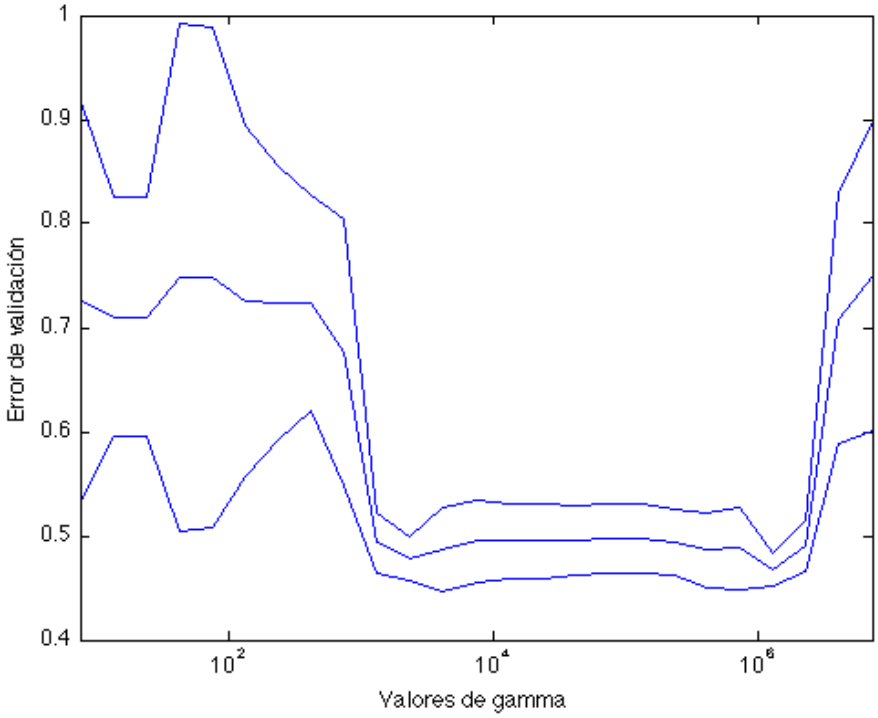

Figura 23. Error de validación del experimento SVM no lineal en función del barrido de Γ

El resultado de esta validación nos permite extraer dos conclusiones. La primera, que podemos reducir el rango de gamma al asignarle valores entre 10^{-1} y 10^3 debido a que la primera parte del rango de gamma no aporta ningún beneficio al experimento. La segunda conclusión es que será necesario validar Γ debido a la inestabilidad del error en función de su valor.

Así pues, en definitiva, las SVM no lineales nos permiten fijar C, por la falta de influencia de su variación en el resultado, y nos exigen que validemos Γ cada vez que sean utilizadas.

- **Resultados**

Por último, llevando a cabo un doble LOO, necesario para la validación, en el que se fija C=100 y se realiza un barrido logarítmico a Γ en función del valor mediano de los datos, el resultado que nos proporciona el experimento global de SVM no lineal empleando todos los datos disponibles es de una tasa de error del sistema de clasificación del 43.6%.

En definitiva, a través del empleo de SVM, tanto lineal como no lineal, las tasas de error resultantes son superiores al 40%, muy próximas al 50% que es el azar. Estos resultados parecen indicar que estamos utilizando mucha información irrelevante que degrada las prestaciones de la máquina, conduciéndonos a la necesidad de utilizar los métodos de selección de características, ya incluidos en el siguiente experimento.

5.2.2. Wrappers

Anterior a los métodos secuenciales, el segundo bloque de métodos globales es el experimento al que hemos denominado 'wrappers'. Se clasifica como método global porque consta de una única fase en la que son utilizados todos los datos, pero por primera vez se incluye la idea de regiones cerebrales y su diferente peso en la detección de TOC. Constituye un punto intermedio entre los métodos globales, en los que no hemos utilizado el concepto de áreas, y los métodos secuenciales, en los que los experimentos se basan en diversas fases en las cuales se trabaja con la información fraccionada según dichas áreas para poder caracterizar la patología.

El experimento al que hemos llamado 'wrappers' se basa en los principios del método de selección de características con el mismo nombre: generar diversos subconjuntos de características y evaluar el predictor con ellos, eligiendo aquel subconjunto que mejores prestaciones aporte.

Tal y como se indicaba en el capítulo anterior, la clave de este tipo de algoritmos se encuentra en elegir cómo construir los diferentes subconjuntos de atributos, punto en el que nosotros vamos a introducir el concepto de regiones cerebrales. Así pues, este experimento se va a basar en utilizar el predictor con todos los voxels del cerebro excepto con aquellos pertenecientes a un área, es decir, emplear en la predicción la información relativa a todas las regiones menos una. Este proceso, llevado a cabo

tantas veces como regiones cerebrales consideramos, permite obtener un error de predicción para cada subconjunto formado extrayendo un área, medida que nos permite valorar la influencia de esa región en el problema de predicción.

En definitiva, el subconjunto que consideraremos óptimo será aquel cuya tasa de error de clasificación haya sido mínima porque ese hecho supone que para generar ese subconjunto se ha extraído la región cuya información resulta más perjudicial en nuestro problema de detección de TOC, ya sea porque sus datos no son discriminatorios o porque contiene información redundante.

Hasta ahí, considerando todos los voxels de un área como una característica conjunta, la construcción de subconjuntos consistiría en extraer una característica por iteración y valorar en qué iteración conseguimos mejores resultados, pero para encontrar un subconjunto óptimo debemos llegar más allá. El siguiente paso es realizar ese bucle de extracción de áreas iterativamente hasta que el subconjunto de características seleccionado esté formado por los voxels de una sola región cerebral. Es decir, si sobre el primer subconjunto óptimo seleccionado, compuesto por todas las áreas menos una, repetimos el proceso de extracción unitaria de regiones para encontrar la segunda región que menos aporta en la resolución de nuestro problema, obtendríamos un subconjunto óptimo formado por todas las áreas menos dos; y si ese proceso lo repetimos tantas veces como áreas tenemos lo que conseguimos es quedarnos únicamente con un área. El esquema del experimento completo sería el siguiente:

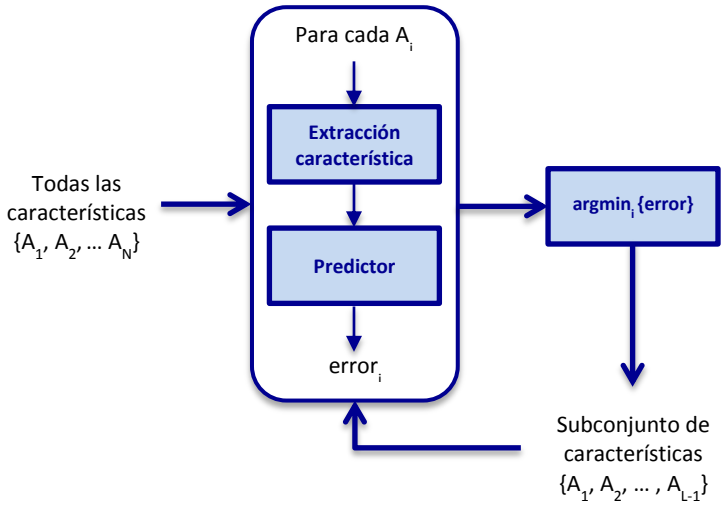

Figura 24. Esquema de funcionamiento de experimento 'wrapper'

Teniendo en cuenta que nuestros datos son divididos en 116 regiones cerebrales, el sistema propuesto nos permite obtener 116 tasas de error asociadas a 116

subconjuntos óptimos de atributos, en el que cada uno tendrá un número de regiones consideradas diferente, desde sólo un área hasta el total de las mismas.

Antes de mostrar los resultados de este experimento, consideramos importante citar que utilizando el mismo esquema se han empleado los dos tipos de predictores que conocemos: una SVM lineal y una SVM no lineal.

- **Resultados**

En el experimento anterior el resultado solo nos proporcionaba una idea de cómo de efectivo era el sistema en la detección de TOC. A diferencia de ése, el sistema denominado como 'wrapper' no solo nos permite estimar la probabilidad de detección de la enfermedad a través de su empleo sino que supone una primera aproximación a la caracterización de la enfermedad, pues la tasa de error mínima obtenida dará a conocer la información de qué áreas, en conjunto, resulta óptima para la resolución de nuestro problema.

Para poder entender el análisis de los resultados de este experimento es necesario tener claro dos ideas:

- Para cada iteración del bucle LOO, en el que un dato se utiliza como test y el resto como subconjunto de entrenamiento, obtendremos un error de validación en relación al número de áreas utilizadas en la predicción. Esa tasa de error es la que vamos a emplear para elegir en cada caso cuántas áreas han sido utilizadas en el caso óptimo de cada iteración.
- De manera global, tendremos una tasa de error del sistema que varíe en función del número de áreas utilizado que nos servirá para definir el error global del sistema propuesto.

Así pues, en primer lugar se muestra, en media, el error de validación obtenido de las 172 iteraciones que lleva a cabo el LOO.

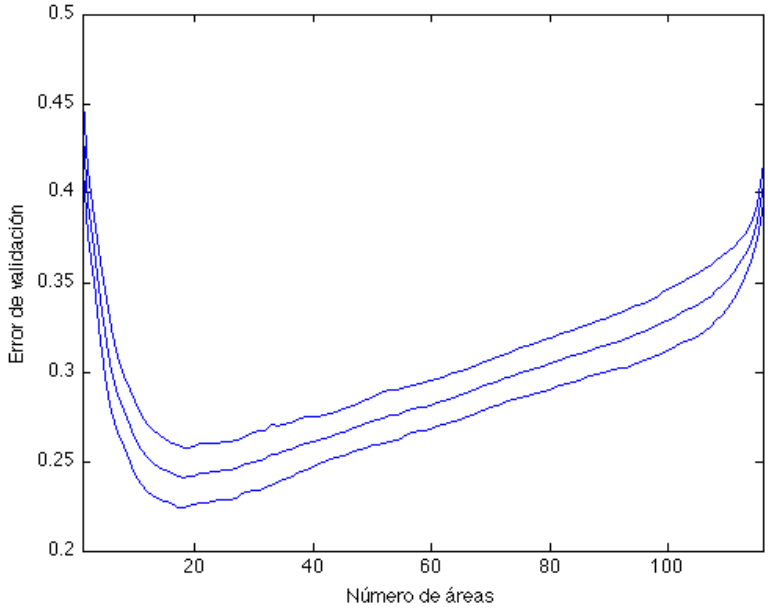

Figura 25. Error de validación del experimento 'wrapper' con SVM lineal en función del número de áreas

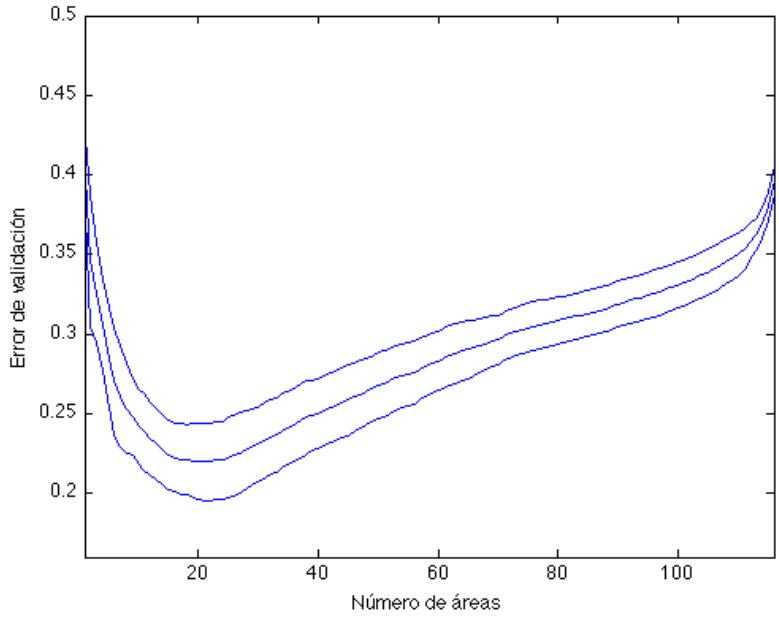

Figura 26. Error de validación del experimento 'wrapper' con SVM no lineal en función del número de áreas

Como podemos observar, en ambos casos, la curva que describe el error tiene siempre una forma similar: con máximos cuando se utilizan demasiadas regiones cerebrales, debido al ruido y a la redundancia que implican algunas áreas, y con los mínimos entorno al uso de 10-30 áreas. Además, podemos ver que cuando el número de áreas se ve muy reducido el error aumenta debido a

que estaríamos eliminando regiones con información relevante para el objetivo que perseguimos.

Por otra parte, utilizando el subconjunto óptimo de cada número de áreas para evaluar el elemento de test obtenemos la tasas de error del sistema, que se representan a continuación.

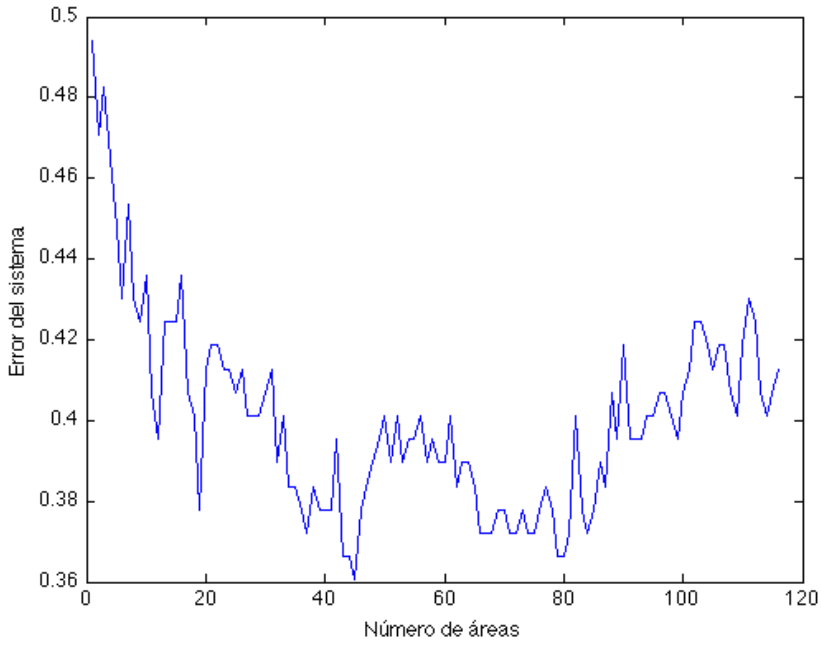

Figura 27. Tasa de error del experimento 'wrapper' con SVM lineal en función del número de áreas

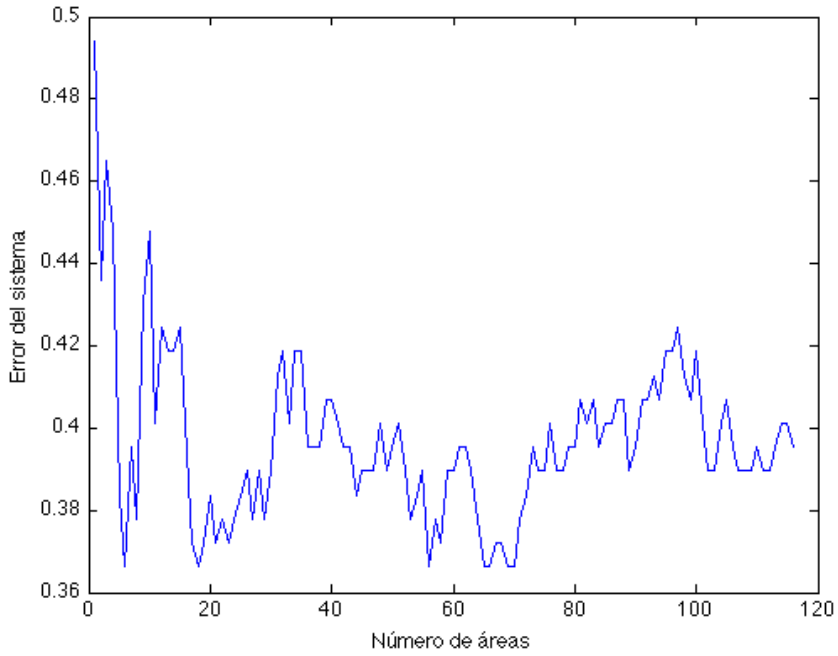

Figura 28. Tasa de error del experimento 'wrapper' con SVM no lineal en función del número de áreas

Por último, vamos a utilizar la información tanto de validación como de predicción del elemento de test para calcular una tasa de error que caracterice al sistema. Para ello, debemos seleccionar el punto de validación óptimo para cada curva en las dos primeras figuras (25 y 26 en cada caso), obteniendo así un subconjunto óptimo diferente para cada iteración del bucle en el que se toma un elemento distinto como test. Así pues, la tasa de error final será el resultado de hacer la media entre las predicciones de test que se han realizado para cada uno de los casos con su subconjunto óptimo particular.

Tras realizar ese cálculo, en la siguiente tabla se recoge la tasa de error que caracteriza cada uno de nuestros sistemas junto a la media de número de áreas empleado para el subconjunto óptimo encontrado en cada caso.

Wrapper	Tasa de error del sistema	Número de áreas medio
Lineal	44.77%	21
No lineal	40.12%	21

Tabla 2. Caracterización, a través de la tasa de error del sistema, del experimento 'wrappers'

Como podemos observar, las tasas de error resultantes son similares a las del apartado anterior, en el que utilizábamos un esquema global con SVMs. Así pues, aunque a través de la selección de características no estemos consiguiendo mejorar las prestaciones de nuestro sistema, sí ganamos interpretabilidad, puesto que hemos eliminado alrededor de 100 regiones que no aportaban ningún tipo de información. Aún así, debido a que cabe esperar que las prestaciones del sistema con la selección de características también mejoren, el análisis de la gráfica del error de test en media en función del número de áreas sugiere que podríamos reducir la tasa de error por debajo del 40% con ese grupo reducido de 21 áreas, resultado con el cual mejoraríamos en interpretabilidad y en clasificación. De esta situación podemos concluir que la validación del número de áreas debe ser mejorada, puesto que el proceso a través del doble LOO no resulta lo suficiente robusto.

5.3. Métodos secuenciales

La segunda categoría en la que se organizan los experimentos desarrollados es la de métodos secuenciales. En ésta, como su propio nombre indica, se encuadrarán todos aquellos experimentos que consten de diferentes fases en cuanto al tratamiento de datos y a la aplicación de técnicas se refiere.

Esta forma de hacerlo por etapas pasa por intentar solucionar el principal problema que surge de los datos de los que disponemos: el escaso número de datos respecto al elevado número de dimensiones de los mismos. Así pues, para resolver dicho inconveniente dividimos los datos de entrada en unidades de menor dimensión, mediante la agrupación de los voxels según las áreas funcionales, y les aplicamos por separado las técnicas de aprendizaje máquina, dando lugar a la primera etapa y surgiendo la necesidad de una segunda etapa en la que los resultados se reúnan y puedan ser analizados de manera conjunta.

A continuación, la Figura 29 contiene una representación del esquema general que emplea el sistema secuencial.

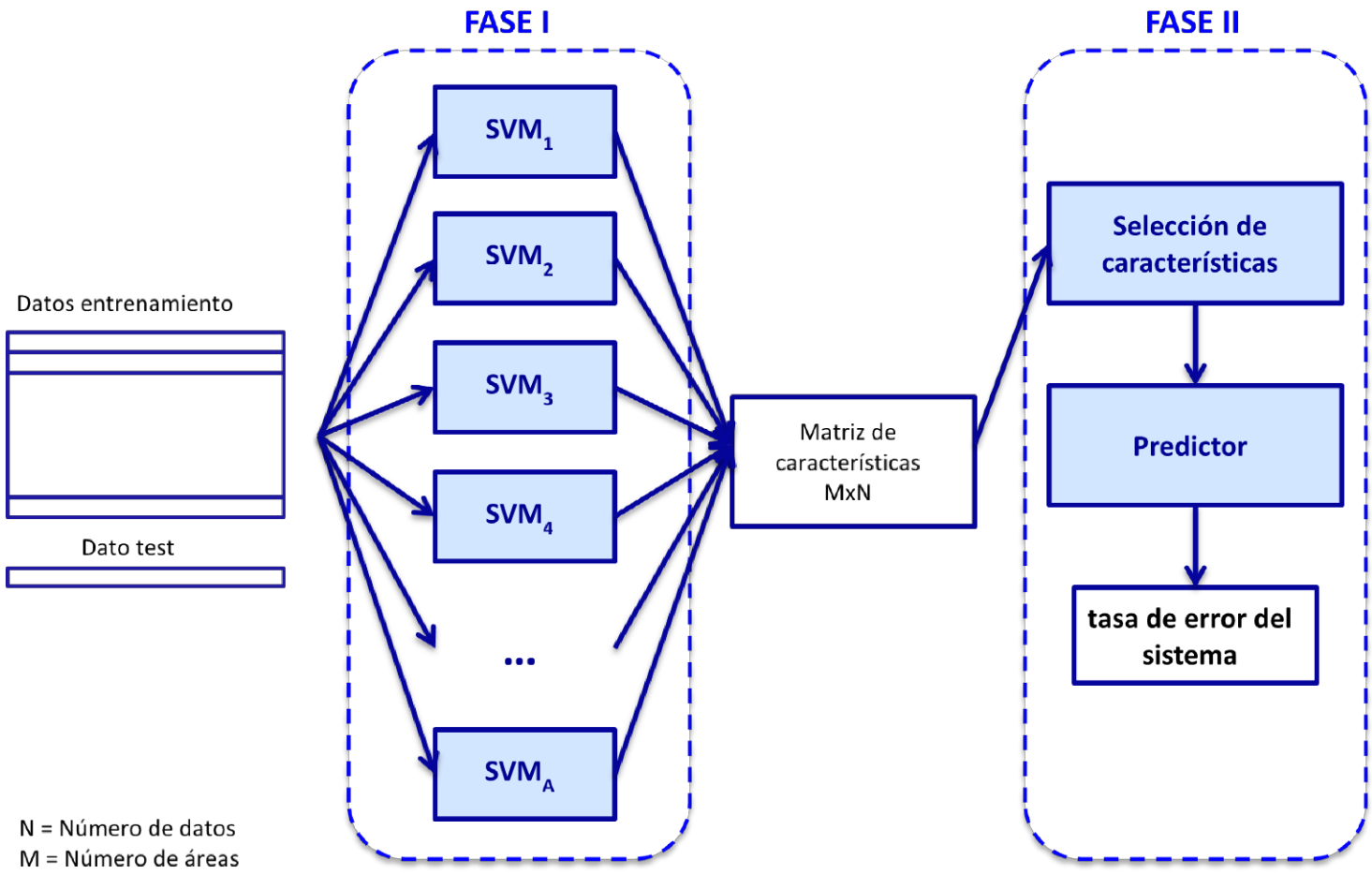

Figura 29. Diagrama de bloques del método secuencial

5.3.1. Fase I: SVM por área

El primer paso a llevar a cabo, antes de analizar qué objetivo persigue esta primera etapa del método secuencial, es descomponer los datos de entrada en varios conjuntos para ser tratados por separado. Esa división de voxels se hace siguiendo la distribución de regiones cerebrales que viene dada según la matriz AREAS_VOXELS definida en la Sección 5.1.1. De esta manera, obtendremos M subconjuntos diferentes, siendo M el número de regiones cerebrales, cuyo contenido, de manera individual, corresponderá a los voxels pertenecientes a un área cerebral.

Como indicábamos anteriormente, el principal objetivo de esta fase es resolver uno de los mayores inconvenientes que se plantean al utilizar imágenes MRI: la alta dimensionalidad de los datos. Así pues, una vez conformados los datos, cada uno de esos subconjuntos serán empleados en la Fase I para caracterizar el área a la que representan a través de un único dato, consiguiendo así reducir drásticamente la dimensionalidad de los datos con los que trabajamos, disminuyendo el coste computacional y facilitando las operaciones posteriores.

La Fase I está basada en que los datos de cada región cerebral pasen por un proceso de aprendizaje máquina. Como el propio título del apartado indica, esta primera etapa de clasificación utilizará M clasificadores SVM, uno por cada área, aplicando cada uno de ellos a un subconjunto diferente de los obtenidos en la división anterior. De esta manera, finalizado el proceso de aprendizaje, por cada clasificador obtendremos una salida blanda que nos servirá para caracterizar cada una de las regiones en las que se encuentran divididos los datos cerebrales.

Por otro lado, como ya ocurría en los métodos globales, la elección a priori del tipo de clasificador SVM es compleja. Por ello, a lo largo de la Fase I de nuestro experimento se va a utilizar tanto el clasificador de tipo lineal como el no lineal.

A través del entrenamiento de una SVM podemos calcular el vector normal al hiperplano de separación entre clases, **w**, por cada uno de los clasificadores del sistema. Así pues, cada SVM nos permite obtener un vector de longitud igual al número de individuos. Esos vectores, al ser reorganizados, pueden emplearse como conjunto de datos de entrada para la selección de características posterior. Esa redistribución queda gráficamente aclarada en la siguiente figura.

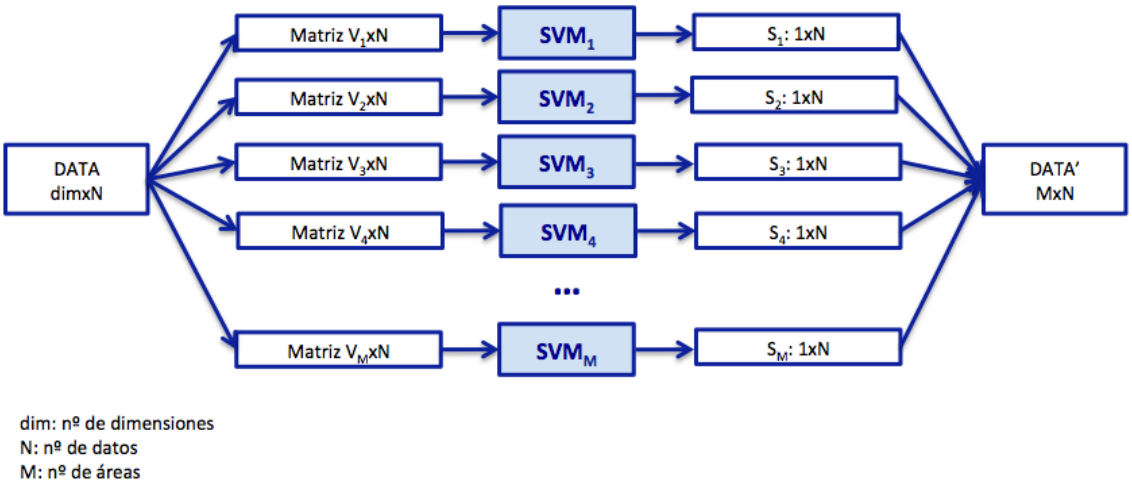

Figura 30. Esquema del funcionamiento de la Fase I, 'SVM por área', del experimento secuencial

- **Generalización del modelo**

Una de las claves de los modelos predictivos es que sean capaces de generar predicciones precisas a partir de la información oculta en los datos. Para ello lo que se busca es generalizar el modelo con los datos disponibles de entrenamiento para que cualquier nuevo dato a evaluar sea clasificado con éxito. En otras palabras, uno de los pasos necesarios de cualquier modelo predictivo es asegurar que no haya sobreajuste de los datos de entrenamiento, ya que esto conduciría a predicciones sub-óptimas. El sobreajuste ocurre cuando un sistema se sobreentrena, pues el aprendizaje queda ajustado a unas características específicas que no tienen relación causal con el objetivo final.

Para generalizar el modelo en esta primera etapa, ésta estará integrada en un LOO que valide el sistema. El fin principal del bucle es calcular la salida blanda para cada individuo a través de una máquina entrenada por todos los datos excepto por si mismo, evitando así que la máquina se ajuste a sus características y provocando que los resultados no sean una caracterización real del experimento.

Por otra parte, el LOO de validación también sirve para calcular un error de validación que nos permita evaluar cómo de bueno está siendo el tipo de SVM utilizado y, además, en el caso de la SVM no lineal el LOO nos servirá para validar el parámetro gamma.

- **Resultados**

Llegados a este punto, a través de esta primera fase no podemos obtener todavía resultados de lo efectivo que es el experimento para la detección de

TOC, pero sí nos permite valorar qué influencia tiene cada tipo de SVM, a través de los errores de validación calculados mediante el LOO.

SVM	Tipo de valor	Error de validación	Área
Lineal	Error mínimo	33.98%	96
Lineal	Error máximo	59.9%	79
No lineal	Error mínimo	50.1%	114
No lineal	Error máximo	64.3%	81

Tabla 3. Errores de validación de la Fase I, SVM por áreas, según el tipo de SVM utilizado

La tabla anterior resume, según el tipo de SVM utilizado, qué valores de validación mínimo y máximo se consiguen de media y para qué áreas. Como podemos observar, la diferencia entre el modelo lineal y el no lineal es significativa, con una diferencia, en media, de un 10-15%. Esta situación nos lleva a que la segunda fase sean utilizados únicamente los datos extraídos a través del uso de la SVM lineal en la primera etapa, pues consideramos que los resultados de la SVM no lineal proporcionarían peores prestaciones.

5.3.2. Fase II: Selección de áreas

La segunda etapa del método secuencial está dividida, como se observa en su esquema de funcionamiento, en dos periodos bien diferenciados. El primero, y más trascendental para el resultado, se basa en los métodos de selección de características con el objetivo de seleccionar aquellas regiones cerebrales cuya información resulte relevante para la detección de TOC y la caracterización de la patología. El segundo se trata de un LOO que entrenará una SVM con los diferentes subconjuntos de áreas seleccionados, lo que nos permite calcular una tasa de error del sistema para evaluar individualmente cada uno de los métodos de selección de características y, de forma global, el sistema secuencial completo.

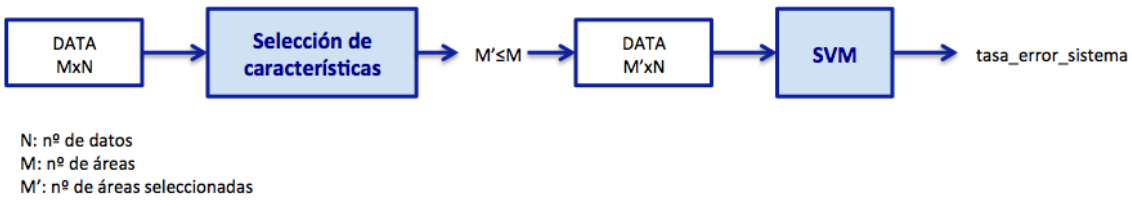

Figura 31. Esquema del funcionamiento de la Fase II, 'Selección de áreas', del experimento secuencial

Así pues, a continuación, en diferentes apartados se desglosan los resultados según las diferentes técnicas de selección de características utilizadas. Por cada una de ellas se comentan las particularidades del método para la aplicación en nuestro

sistema y, posteriormente, se muestran los resultados, analizados a través de dos datos:

-la gráfica que muestra la variación del error según el número de áreas utilizadas.
-la tasa de error de clasificación que caracterice al sistema completo para el uso de ese método de selección de características concreto.

5.3.2.1. Ranking de variables

El objetivo principal del ranking de variables es ordenar los atributos de los que disponemos, en nuestro caso las áreas cerebrales. de mayor a menor relevancia. Ese ranking nos servirá para formar M subconjuntos, siendo M el número de áreas, con los que evaluar la SVM que forma la segunda parte de esta Fase II. En los subconjuntos el primero estará formado por la región más relevante, y para conformar el resto se irán añadiendo de una en una las siguientes áreas en la lista de trascendencia. Al final, lo que se busca es encontrar el número de áreas con el que se optimiza el resultado.

Como ya se indicó en el capítulo anterior, la clave para utilizar este tipo de algoritmo es definir qué criterio se va a utilizar para calcular la medida de relevancia de cada característica. En nuestro caso, derivado de las operaciones de la fase en la que se ha llevado a cabo una SVM por área, tenemos un error de validación por área para cada iteración del bucle LOO donde se separan datos de entrenamiento y elemento de test. Esa tasa de error de validación es usada en nuestro experimento directamente como medida de relevancia de cada región cerebral, considerando que las áreas con menor error de validación son aquellas más relevantes para la resolución de nuestro problema, pues serían aquellas que aportan mayor información sobre la presencia de la patología.

Una vez decidido el criterio de relevancia se evalúa con cada subconjunto elegido tanto una SVM lineal como una SVM no lineal.

- **Resultados**

Primeramente, en las siguientes gráficas podemos encontrar información acerca de la tasa de error de clasificación resultante de utilizar subconjuntos con diferente número de áreas. Éstos son construidos a partir de la lista de regiones cerebrales más relevantes.

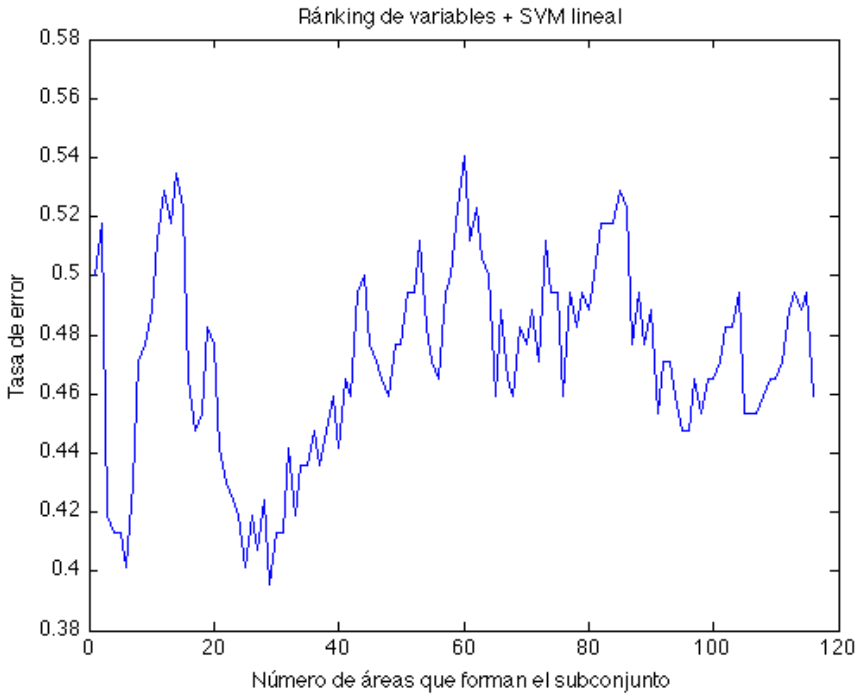

Figura 32. Tasa de error del sistema secuencial con ranking de variables según el número de áreas que forman el subconjunto de regiones elegido. Para la evaluación del sistema es empleada una SVM lineal

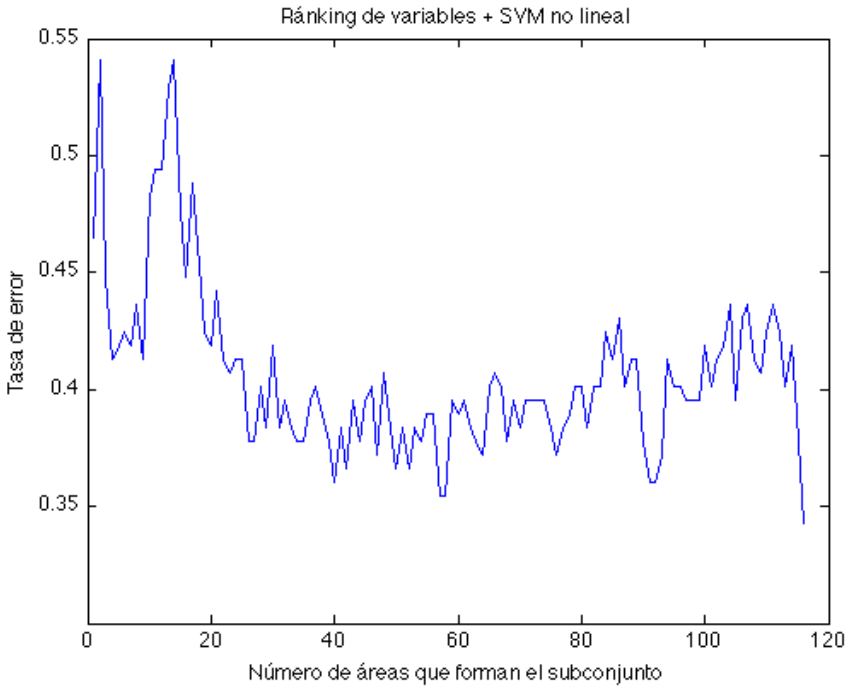

Figura 33. Tasa de error del sistema secuencial con ranking de variables según el número de áreas que forman el subconjunto de regiones elegido. Para la evaluación del sistema es empleada una SVM lineal

En estas figuras podemos ver de forma clara la diferencia en resultados que supone usar un tipo de SVM u otro, ya que encuentran sus mínimos y máximos

en zonas completamente diferentes y el perfil de la tasa de error también es totalmente distinto.

Por otra parte, la forma de caracterizar el sistema analizar los datos la tasa de error de clasificación mínima. En la siguiente tabla, distinguiendo entre el uso de una SVM lineal y una no lineal, se da a conocer el error mínimo que conseguido y cuántas áreas tiene el subconjunto de regiones con el que se obtiene dicho error.

SVM	Tasa de error mínima	Número de áreas
Lineal	39.53%	29
No lineal	34.3%	116

Tabla 4. Resumen de la tasa de error del sistema secuencial utilizando ranking de variables como método de selección de características en la segunda fase

Por un lado estos datos nos permiten saber que, con un subconjunto de áreas mucho menor que el total, el SVM lineal es capaz de minimizar su tasa de error en la detección de TOC, lo que confirma nuestra teoría inicial de que para la clasificación muchas áreas suponen información redundante o irrelevante. Por otro lado, la SVM no lineal necesita de todas las áreas disponibles para hacer el mejor de sus diagnósticos, lo que conlleva que la selección de características a través del ranking de variables no aporte nada si la predicción final va a ser llevada a cabo mediante una SVM no lineal. Aún así, en la gráfica en la que se representa la tasa de error en función del número de áreas podemos observar que existe un punto entorno al empleo de 60 áreas en el que las prestaciones serían similares a utilizar 116, es decir, un error entorno al 35%, ganando en interpretabilidad por la eliminación de más de 50 áreas. Este punto representa un compromiso intermedio entre la mejora de prestaciones y la búsqueda de interpretabilidad.

5.3.2.2. Wrappers

El segundo de los métodos de selección de características utilizado en nuestro sistema es un 'wrapper'. Esta técnica se basa en formar subconjuntos de características y evaluar un predictor con ellos, eligiendo como óptimo aquel subconjunto que aporte mejores prestaciones en la detección de TOC y caracterización de la patología.

Como ya ocurría en su utilización como método global, la esencia de su empleo reside en seleccionar los subconjuntos con los que evaluar el predictor.

El criterio para ello es el mismo que en el caso anterior: generar tantos subconjuntos como áreas cerebrales tenemos extrayendo en cada uno la información de una de las regiones. De esta manera, la tasa de error de clasificación menor implicará que el subconjunto con el cual se obtiene ha sido formado extrayendo la región cuya información influye en menor medida en nuestro problema de predicción. Además, de nuevo como en el caso del método global que sigue el mismo principio, ese bucle se repetirá hasta que el subconjunto de áreas elegido esté formado por solo un área, es decir, en cada iteración elegiremos un subconjunto formado por un área menos, lo que nos permite construir la gráfica de error de validación en función del número de áreas utilizado en el subconjunto óptimo.

Por último, según el tipo de SVM, lineal o no lineal, elegido en el predictor que permite evaluar qué subconjunto es el óptimo, diferenciamos el 'wrapper' en dos categorías: 'wrapper' lineal y 'wrapper' no lineal.

- **Resultados**

En primer lugar, según el número de áreas que forman el subconjunto óptimo, se halla un error de clasificación del sistema completo representado en las siguiente gráficas para cada uno de los tipos de 'wrapper', lineal y no lineal, que se han utilizado.

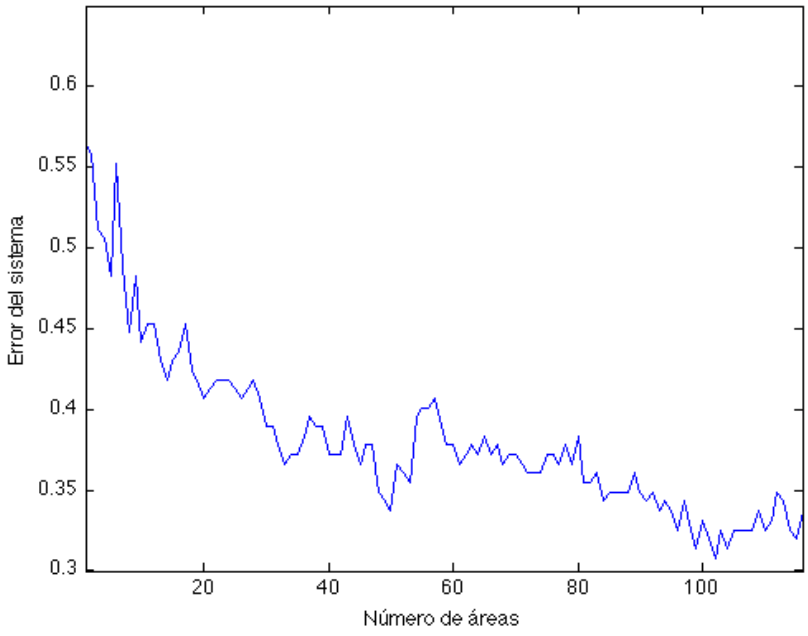

Figura 34. Tasa de error de clasificación del método global empleando como selección de características un 'wrapper' lineal

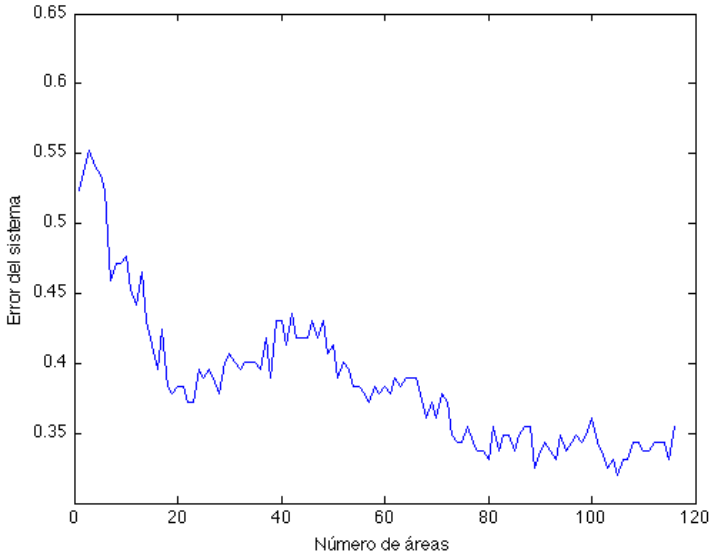

Figura 35. Tasa de error de clasificación del método global empleando como selección de características un 'wrapper' no lineal

Por otra parte, debemos tener en cuenta que para hallar una tasa de error de clasificación que caracterice al sistema ha sido necesario utilizar un LOO. En él cada iteración lleva a cabo otro LOO, dando lugar al doble LOO, para validar el subconjunto óptimo. Estas circunstancias nos proporcionan un subconjunto óptimo diferente para cada iteración del LOO principal, en que se utiliza un dato como test y el resto como subconjunto de entrenamiento. Así pues, las siguientes gráficas a analizar son las que nos muestran, en media, el error de validación de las 172 iteraciones del LOO según el número de áreas del subconjunto óptimo.

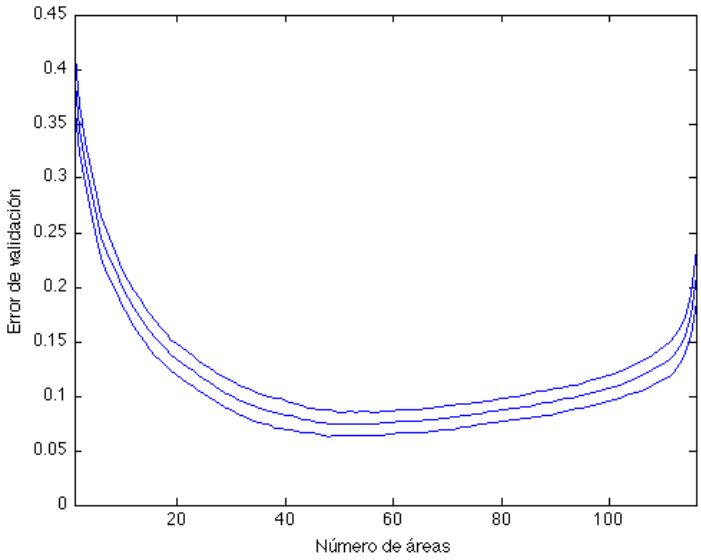

Figura 36. Error de validación del método secuencial, empleando 'wrapper' lineal como selección de características, en función del número de áreas utilizado

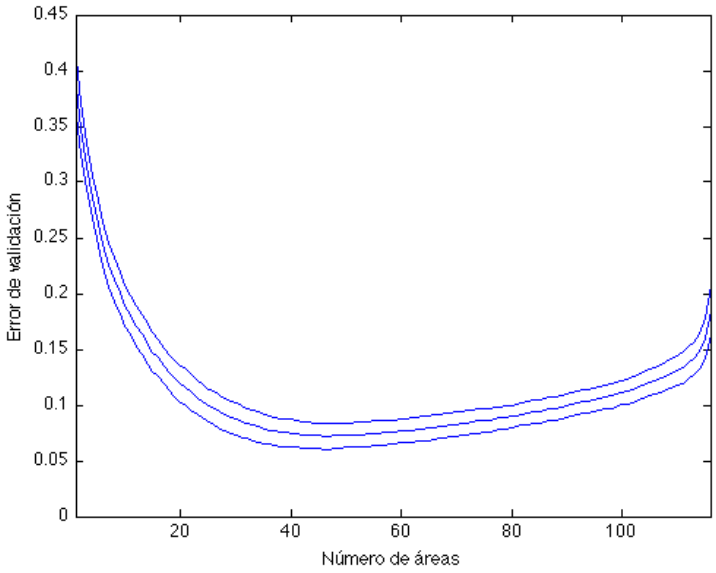

Figura 37. Error de validación del método secuencial, empleando 'wrapper' no lineal como selección de características, en función del número de áreas utilizado

Como en las anteriores figuras podemos observar, a través del empleo del 'wrapper' como método de selección de características en el método secuencial, obtenemos un resultado con implicaciones similares que al utilizar la técnica como método global: la tasa de error se minimiza al utilizar entre 20 y 50 áreas, lo que nos permite concluir que es ese rango de cantidad de áreas donde se encuentra la información relevante para detectar pacientes de TOC y caracterizar la patología, puesto que el resto de áreas suponen información redundante.

Analizando los últimos resultados de validación, si localizamos en las figuras la posición donde la tasa de error de clasificación es mínima conseguimos conocer el número de áreas que forman el subconjunto óptimo para cada iteración del LOO. A partir de ese punto óptimo se utiliza la predicción del elemento de test en esa iteración para calcular la tasa de error total.

De este modo, la caracterización total del sistema secuencial utilizando 'wrappers', a través de la tasa de error de clasificación y del número de áreas medio que forman el subconjunto óptimo, es la siguiente:

Wrapper	Tasa de error mínima	Número de áreas
Lineal	37.21%	56
No lineal	40.7%	51

Tabla 5. Resumen de la tasa de error del sistema secuencial utilizando 'wrappers' como método de selección de características en la segunda fase

5.3.2.3. Eliminación recursiva de características

Como siguiente técnica de selección de características hemos empleado el primero de los métodos de tipo 'embedded' utilizados: la eliminación recursiva de características. En concreto, en nuestro caso se trata de una eliminación hacia atrás, es decir, partimos de todo el conjunto de características y vamos eliminando una a una aquella menos relevante. Esa extracción de características de una en una nos permitirá utilizar la misma representación de la tasa de error de clasificación que en los casos anteriores, en función del número de áreas empleado.

Cabe recordar que ésta es una técnica relacionada directamente con la utilización de SVM, pues emplea la salida de entrenamiento de la SVM para conocer qué característica provoca mayor reducción del margen que separa las clases con las que estamos tratando. Esa estrecha relación con las SVM conlleva que este método se subdivida en dos técnicas a aplicar por separado, según si utilizamos una SVM lineal o una no lineal para entrenar la máquina que evalúa la influencia de cada característica en el margen. Incidimos en esa puntualización puesto que para conocer el peso de cada característica el proceso es diferente según el tipo de SVM que utilicemos.

El objetivo final de la técnicas utilizando un tipo u otro de SVM es el mismo, evaluar iterativamente la extracción de qué área provoca la mínima variación del margen. Al emplear el kernel lineal tendremos acceso a todos los valores del vector de pesos **w**, por lo que es suficiente con conocer cuál de sus componentes tiene menor valor absoluto, pues será aquella que menos esté aportando en la maximización del margen. Sin embargo, para el caso no lineal, el proceso es más complejo porque para conocer la influencia de cada área en el margen es necesario recalcular la matriz de kernel para poder obtener la norma 2 del vector de pesos **w**. En este caso se extrae la componente en la que la diferencia entre la norma 2 utilizando todas las regiones y la norma 2 habiendo extraído dicha componente sea mínima.

- **Resultados**

Así pues, a continuación se representan los resultados de aplicar esta técnica con sus dos vertientes como selección de características sobre los datos extraídos de la Fase I.

Primeramente, como en los casos anteriores, en las siguientes figuras podemos observar cómo existe un rango de número de áreas en el que se encuentran los valores mínimos de la tasa de error de clasificación. De nuevo esto es debido a la existencia de áreas cerebrales que aportan redundancia o información que puede ser considerada como ruido en la búsqueda de la caracterización del TOC.

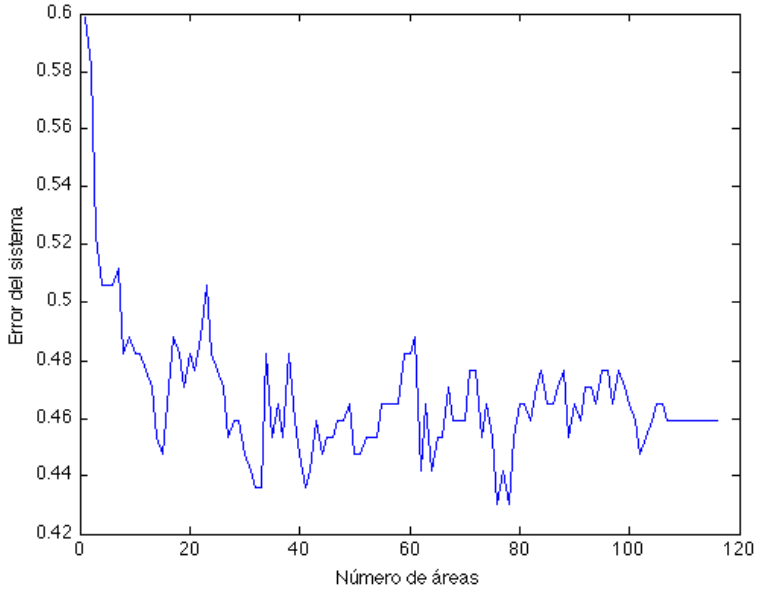

Figura 38. Error de clasificación del método secuencial, empleando RFE lineal como método de selección de características, en función del número de áreas utilizado

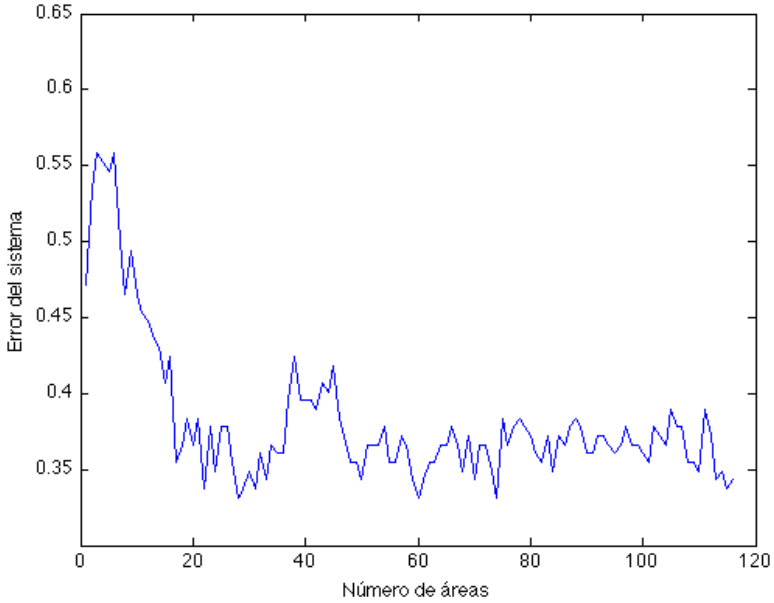

Figura 39. Error de clasificación del método secuencial, empleando RFE no lineal como método de selección de características, en función del número de áreas utilizado

En definitiva, para caracterizar cada uno de los sistemas utilizamos el valor del error de clasificación mínimo, teniendo en cuenta como información complementario qué cantidad de áreas cerebrales es empleada en el subconjunto que da lugar a esa tasa mínima.

RFE	Tasa de error mínima	Número de áreas
Lineal	43.02%	39
No lineal	33.14%	43

Tabla 6. Resumen de la tasa de error del sistema secuencial utilizando RFE como método de selección de características en la segunda fase

En este caso, hemos de tener en cuenta que, por limitaciones de cómputo, no se ha empleado la validación cruzada que lleva a cabo el doble LOO necesaria para encontrar el mejor punto de trabajo. A pesar de ello, el error de test podemos ya muestra indicios del potencial de este método, principalmente en su versión no lineal, puesto que proporciona tasas de error del 33% con 43 áreas. Así confirmamos nuestra teoría de que a través de la selección de características mejoramos el método global en prestaciones e interpretabilidad.

5.3.2.4. SVM norma 1

El último de los métodos de selección de características empleado en la segunda fase del método secuencial se trata del modelo SVM norma 1.

Como recordatorio, cabe destacar que este método sustituye al algoritmo SVM estándar a través de una optimización lineal que permite minimizar la norma 1 del vector de pesos **w**. Dicha optimización se lleva a cabo mediante la regularización de parámetros del modelo, lo que permite conseguir una solución dispersa.

En el Apartado 4.2.4. de este documento se encuentra la formulación que sigue esta técnica para obtener la discontinuidad en el origen que da lugar al truncado de los coeficientes cercanos a 0, es decir, la formulación que nos proporciona la selección automática de características. Dicha selección será más o menos amplia dependiendo del parámetro de regularización que empleemos.

Así pues, al emplear este modelo en nuestro sistema la decisión más importante es elegir el rango que vamos a barrer para darle valor al parámetro C, el cual regular la importancia que le damos a los errores en las SVM.

- **Resultados**

Como con otros métodos anteriormente, para empezar a utilizar el modelo SVM norma 1 hay que llevar a cabo una elección: el rango para asignar valores al parámetro C. En esta ocasión el tamaño de los subconjuntos de características elegidos no será regulado directamente por nosotros sino que según el parámetro C el truncado de los coeficientes cercanos a 0 será mayor o menor, conllevando un tamaño variable de dichos subconjuntos.

Estas circunstancias dan lugar a que varios valores de los asignados a C puedan dar lugar a un subconjunto de igual tamaño, independientemente de que las características seleccionadas sean o no iguales. Dado que estamos analizando el impacto que tienen el número de áreas elegidas en la detección de la patología, en el caso de que varios C den lugar a subconjuntos del mismo tamaño evaluaremos el sistema promediando su tasa de error de clasificación.

Así pues, como primer contacto hemos utilizado un rango conformado por un barrido logarítmico de 100 valores que comprende entre 10^{-2} y 10^{5}. En la siguiente figura se representa en media el número de áreas que componen el subconjunto óptimo en función de los valores de C.

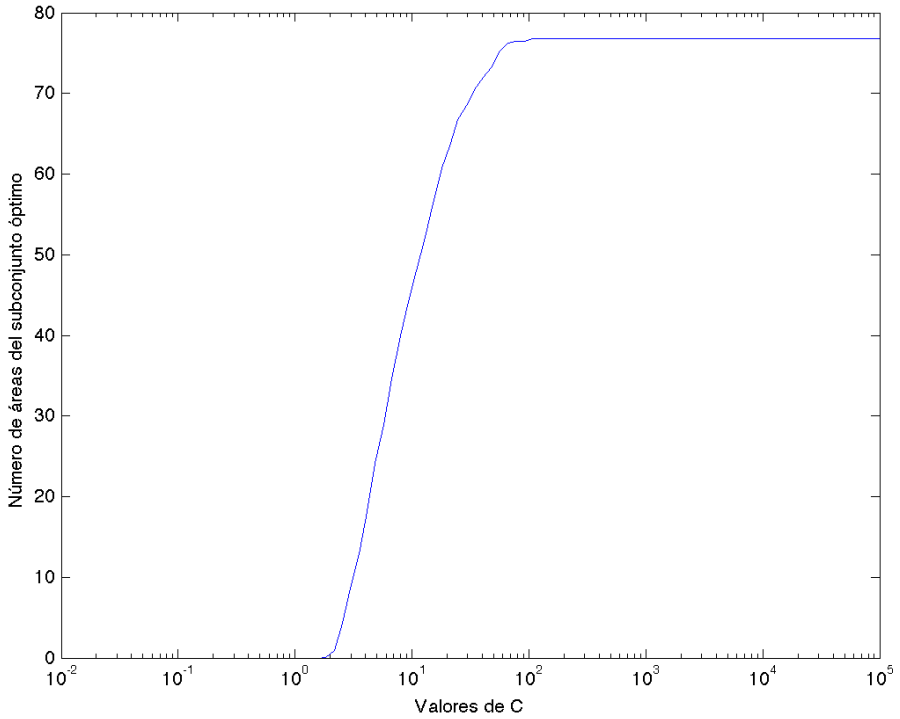

Figura 40. Número de áreas que conforman el subconjunto óptimo según la selección automática de variables del modelo SVM norma 1 en función del parámetro de libertad C

Esta representación nos permite reducir el rango de C a un rango entre 1 y 10^2, debido a que para valores más pequeños al primero no existe subconjunto óptimo porque todos los coeficientes son truncados a 0, y para valores mayores al segundo el número de áreas que conforman el subconjunto óptimo siempre es 77. De esta particularidad concluimos que el método SVM norma 1 siempre considera entre 35 y 40 áreas irrelevantes para nuestro objetivo: detectar los pacientes de TOC y caracterizar la patología.

Finalmente, empleando un rango de 1000 valores repartidos logarítmicamente entre 1 y 10^2 la tasa de error del sistema se representa en la siguiente figura.

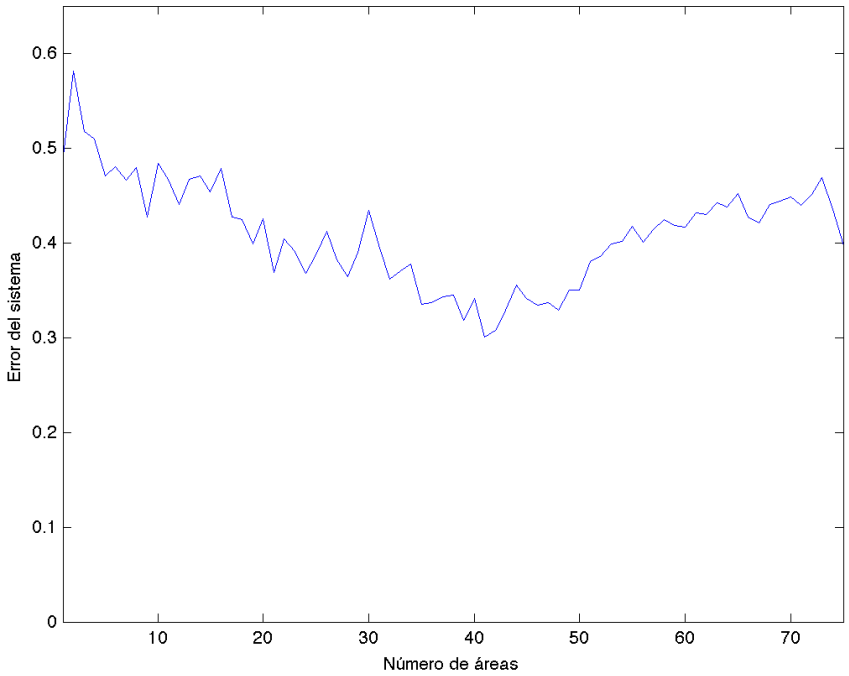

Figura 41. Error de clasificación del método secuencial, empleando SVM norma 1 como método de selección de características, en función del número de áreas utilizado

En primer lugar, en este caso debemos tener en cuenta que sólo se representa hasta el empleo de 75 áreas porque la selección automática realizada a través del modelo SVM norma 1 selecciona como máximo entorno a ese número de áreas, generando una discontinuidad.

En la representación podemos observar que ocurre como con métodos anteriores: al utilizar pocas áreas el error es mayor debido a que algunas de las áreas extraídas aportan información concluyente acerca de la patología,

mientras que al utilizar muchas áreas, la información redundante también provoca que la tasa de error de clasificación se eleve. Así pues, el punto óptimo que caracteriza al método secuencial utilizando el modelo SVM norma 1 para seleccionar las características es aquel en el que se utilizan 41 áreas, siendo la tasa de error de clasificación relacionada del 30.11%.

Aún así, hemos de tener en cuenta, como en el caso anterior, que por limitaciones de cómputo no se ha llevado a cabo el doble LOO que nos permita validar el sistema, por lo que los datos no son la caracterización real del experimento, pero sí nos sirven como guía para conocer el potencial del modelo SVM norma 1 como método de selección de características.

5.4. Análisis comparativo de los diferentes métodos

Para terminar con el capítulo de desglose de los experimentos con sus correspondientes resultados, las siguientes tablas resumen, a través de datos numéricos, el rendimiento de los diferentes algoritmos desarrollados.

La primera de ellas recoge la tasa de error de clasificación de aquellos experimentos que, al haber sido validados, podemos considerar caracterizados a través de ese dato, además del número óptimo de áreas que utiliza para caracterizar la patología. Por otro lado, en la segunda tabla podemos encontrar el mínimo error de clasificación que todos los métodos han obtenido, vinculado al número de áreas que componen el subconjunto con el que se obtiene dicho error.

Método	Selección de características	Clasificación	Error de clasificación	Nº óptimo de áreas
Métodos globales	-	SVM lineal	41.28%	-
		SVM no lineal	43.6%	-
	Wrapper	SVM lineal	44.77%	22
		SVM no lineal	40.12%	22
Métodos secuenciales	Wrapper	SVM lineal	37.21%	56
		SVM no lineal	40.7%	51

Tabla 7. Cuadro resumen de resultados de los algoritmos desarrollados para diagnosticar y caracterizar TOC tras su validación

Método	Selección de características	Clasificación	Error de clasificación	Nº óptimo de áreas
Métodos globales	-	SVM lineal	41.28%	-
		SVM no lineal	43.6%	-
	Wrapper	SVM lineal	36.05%	44
		SVM no lineal	36.63%	70
Métodos secuenciales	Ranking de variables	SVM lineal	39.53%	29
		SVM no lineal	34.3%	116
	Wrapper	SVM lineal	30.81%	102
		SVM no lineal	31.98%	105
	RFE lineal	SVM lineal	43.02%	39
	RFE no lineal	SVM no lineal	33.14%	43
	SVM norma 1	SVM lineal	30.11%	41

Tabla 8. Cuadro resumen de las tasas de error de clasificación mínimas obtenidas con los algoritmos desarrollados para diagnosticar y caracterizar TOC

Como podemos observar, el error de test mínimo siempre es menor que la tasa de error validada que nos permite caracterizar el sistema. Sin embargo, aunque el procedimiento justo para seleccionar el número de áreas óptimo sea la validación cruzada, la tasa de error mínima es suficiente para qué métodos son más prometedores y, en ese caso, validar el experimento concreto, puesto que computacionalmente la validación de todos ellos no es viable.

Además, para reducir las diferencias considerables entre los errores tras validar y las mejores tasas de error de clasificación, sería conveniente refinar la validación cruzada a través del empleo de un “boostrap”, método de re-muestreo aleatorio de datos que permite garantizar la independencia de los resultados de la división entre datos de entrenamiento y test.

Capítulo 6

Conclusiones y líneas futuras

En el presente trabajo se han definido una serie de algoritmos a través de los cuales llevar a cabo experimentos con el objetivo de conseguir optimizar la detección y la caracterización de TOC. En este capítulo se revisan los resultados obtenidos, utilizándolos como indicadores de qué objetivos se han alcanzado y posibilitando la extracción de conclusiones. Posteriormente, se proponen algunas líneas futuras de trabajo entorno a este contenido.

6.1. Análisis de resultados

Para poder analizar los resultados es importante tener claro cuál era el objetivo final de este proyecto. La premisa inicial era conseguir la detección y caracterización del trastorno obsesivo compulsivo basándonos en encontrar ciertas diferencias estructurales entre sujetos sanos y pacientes. Así pues, a continuación se enumeran las conclusiones que se podemos extraer de la aplicación de nuestros algoritmos.

- *Estructura cerebral*: Sin utilizar selección de características, es decir, basándonos tanto en los resultados de los primeros métodos globales, SVM lineal y SVM no lineal, y en la primera fase del método secuencial, el empleo de una SVM lineal obtiene ligeras mejoras en los resultados que la SVM no lineal, lo que nos permite deducir que probablemente las relaciones entres los vóxeles del cerebro son lineales.

- *Selección de características*: En primer lugar, su empleo supone una clara ventaja en la caracterización de la patología, puesto que reduce la tasa de error de clasificación, permitiéndonos acotar en unas áreas concretas la información concluyente para la detección de TOC.

- *Métodos de selección de características:* En primer lugar, para compararlos utilizaremos la tabla 8 del apartado anterior, puesto que por limitaciones computacionales no se han validado todos los experimentos. Al comparar los métodos de selección de características podemos concluir que aquellos que mejores prestaciones nos aportan son el modelo SVM norma 1 y la utilización de

un wrapper lineal. En concreto, si de elegir uno se tratase, seleccionaríamos el modelo SVM norma 1 porque, además de que su selección reduce más la tasa de error de clasificación, es decir, aporta mejores prestaciones, al tratarse de un método "embedded" la carga computacional que supone es mucho menor que cualquier wrapper.

A esta conclusión llegamos a través de la interpretación los resultados de los que disponemos como guía; sin embargo, para confirmarlo sería necesario llevar a cabo la validación con doble LOO de la que hemos hablado en apartados anteriores.

- *Número de áreas*: Centrándonos en el objetivo de caracterizar la patología, resulta de vital importancia analizar qué cantidad de áreas resultan relevantes y cuáles aportan información redundante en nuestro problema. Así pues, como en el cuadro resumen podemos observar, existe un grupo formado por entre 30 y 50 áreas que resultan significativas en la detección de TOC, mientras que el resto podrían ser prescindibles. Con esta conclusión se avanza en el objetivo de disminuir la dimensionalidad de los datos de entrada y conseguimos eliminar información redundante que pueda perjudicar nuestro resultado.

6.2. Líneas futuras

A pesar de todo el trabajo realizado, quedan muchos trabajos futuros de investigación relacionados con este campo. Algunas de las posibles líneas a seguir para continuar mejorando la propuesta presentada en este proyecto son:

- Validar todos y cada uno de los experimentos. La validación se ha llevado a cabo de manera individual y por separado de cada uno de los algoritmos, pero podría resultar interesante analizar los resultados de validar cada experimento de manera global. Durante el presente trabajo se descartó esta posibilidad porque se priorizo la inversión de nuestro esfuerzo en la generación de más métodos de selección de características.

  Así pues, como trabajo futuro quedaría incluir en todos y cada uno de los experimentos un doble LOO que nos permita validar el número de áreas seleccionado. Además, en el caso del método global 'Wrapper no lineal', como ya comentábamos acerca de los resultados, la selección de áreas no parece óptima, así que la validación por medio de un doble LOO podría ser mejorada con un "boostrap".

- Caracterizar por completo la patología identificando las regiones que forman el grupo de mayor relevancia a la hora de detectar el TOC. Para llegar a ello el

trabajo técnico llegaría al punto de señalar qué áreas de la división de la que partimos contienen mayor información relevante, mientras que sería necesaria la participación de un neurólogo o un psiquiatra especialista para poder realizar la correspondencia entre esas áreas y su distribución física en el cerebro.

- Comparar la relación existente entre las regiones detectadas por los algoritmos desarrollados con las regiones sugeridas por otros estudios similares en este campo, además de realizar un balance sobre qué aportan estos experimentos respecto a los métodos tradicionales. En este punto se podría incluir la comparación de la caracterización de la patología a través de aprendizaje máquina con el diagnóstico real de un médico especialista.

- Realizar una investigación más meticulosa, lo que conllevaría disponer de más estudios. Esta opción depende del presupuesto, puesto que las imágenes de resonancia implican un gran coste, y de la disponibilidad de pacientes de TOC a los que realizar las pruebas correspondientes.

- Empleo de otro tipo de datos para la predicción, combinados adecuadamente con los que hemos utilizado de MRI estructural. Entre éstos se podrían utilizar variantes de neuroimagen como, por ejemplo, la MRI funcional, información demográfica de los pacientes, información genética, etc. Con ello podríamos relacionar áreas funcionales con implicaciones en la patología con las áreas estructuras que hemos categorizado como relevantes para nuestro objetivo.

Capítulo 7

Presupuesto y planificación

En este capítulo se presenta, por un lado, la planificación y estrategia de seguimiento empleada en la realización del proyecto, y por otro, la justificación de los costes globales que supone la misma.

7.1. Planificación

El inicio del proyecto consistió en una etapa de documentación de duración inicialmente sin estimar en la que los objetivos principales eran la transferencia de conocimientos del tutor al alumno y que el alumno tomara conciencia de las investigaciones existentes en el ámbito del diagnóstico de enfermedades psíquicas a través de la aplicación de procesado digital de imágenes.

Tras esa primera etapa, la organización del trabajo se ha llevado a cabo a través de varias tareas con fines muy diferenciados y mediante revisiones con el tutor que servían para marcar el rumbo a tomar en los siguientes pasos. Además, las dudas y los problemas puntuales han sido resueltos mediante correo electrónico independientemente de dichas reuniones.

A continuación podemos encontrar la descripción gráfica y sintética de la organización a través de dos elementos. El primero, una tabla donde se recoge cada una de las tareas, junto a su fecha de inicio, fecha de fin, tarea predecesora a la que se encuentra vinculada, en el caso de que exista, y una pequeña descripción aclaratoria cuando se considera necesario. El segundo elemento se trata del diagrama de planificación, para el cual se ha utilizado el diagrama de Gantt, herramienta gráfica de programación de proyectos de diversa naturaleza cuyo objetivo es mostrar el tiempo de dedicación previsto para las diferentes tareas.

Nombre de la tarea	Fecha de inicio	Fecha de finalización	Duración	Predecesoras	Comentarios
Fase previa	**02/07/12**	**14/09/12**	**55**		
Planificación	02/07/12	03/07/12	2		
Transferencia de conocimientos	04/07/12	06/07/12	3	2	Puesta en común de objetivos
Estudio de alternativas tecnológicas	09/07/12	17/08/12	30	3	Documentación sobre análisis de imagen, SVMs, selección de características
Estudio de investigaciones existentes	09/07/12	17/08/12	30	3	Documentación sobre estudios previos
Análisis de datos	03/09/12	14/09/12	10		Familiarización con los datos
Algoritmos base	**17/09/12**	**26/10/12**	**30**	**1**	**Desarrollo de la base de los experimentos**
SVM lineal	17/09/12	25/09/12	7		
SVM no lineal	26/09/12	28/09/12	3	8	
Selección de características	01/10/12	12/10/12	10	8, 9	SVM norma 1, RFEs, 'wrappers'
Pruebas de algoritmos	15/10/12	26/10/12	10	8, 9, 10	
Experimentos	**29/10/12**	**17/04/13**	**123**		
Diseño de experimentos	29/10/12	31/10/12	3	7	Boceto de la estructura
Métodos secuenciales: Fase I	01/11/12	28/11/12	20	13	SVM por área
Métodos secuenciales: Fase II	29/11/12	26/12/12	20	13, 14	Selección de áreas relevantes
Métodos globales: SVM	27/12/12	02/01/13	5	13, 14, 15	
Métodos globales: Wrappers	03/01/13	23/01/13	15	13, 16	
Validación y análisis de resultados	**24/01/13**	**17/04/13**	**60**		
Lanzamiento scripts experimentos	24/01/13	20/02/13	20	16, 17	
Validación parámetros	21/02/13	03/04/13	30	19	Optimización de los parámetros de los que dependen los experimentos
Análisis de resultados	04/04/13	17/04/13	10	19, 20	
Memoria	**18/04/13**	**29/05/13**	**30**	**12**	

Tabla 9. Esquema de tareas

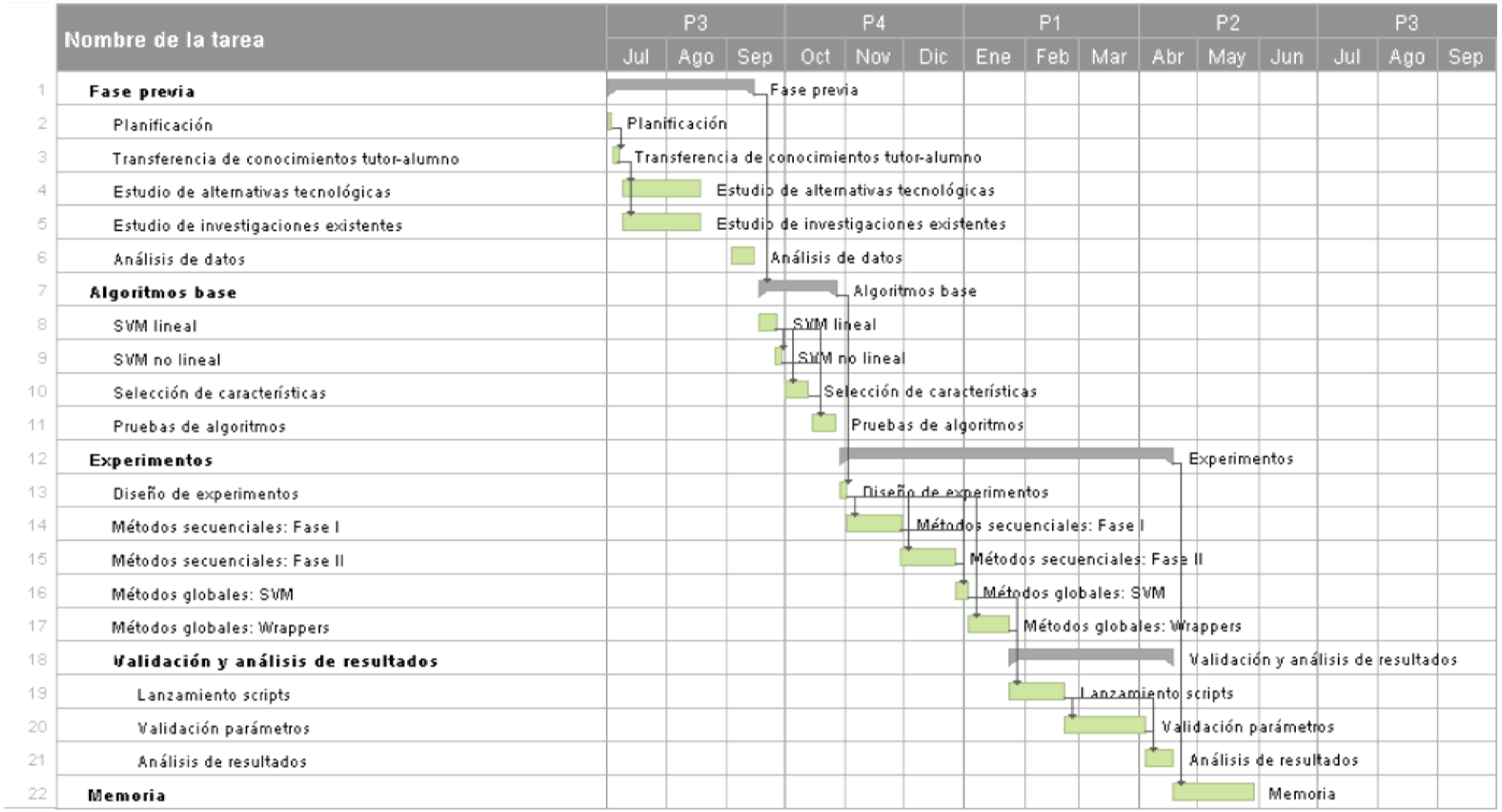

Figura 42. Diagrama de Gantt

7.2. Presupuesto

En este apartado se presentan justificados los costes globales de la realización del proyecto. Tales costes se muestran desglosados según las categorías de personal y material, para posteriormente ofrecer un cuadro resumen con el presupuesto total que supone el proyecto.

7.2.1. Costes de personal

Los costes de recursos humanos relacionados con el proyecto están vinculados directamente con la planificación del mismo. Así, como podemos comprobar en la tabla 10, el desarrollo total del proyecto se divide en 4 fases. La primera corresponde al periodo de documentación y familiarización con el entorno por parte del ingeniero proyectante. Tras ella, es turno de la implementación de los diferentes algoritmos necesarios para los experimentos. En tercer lugar, se lleva a cabo el diseño, ejecución y validación de dichos experimentos. Y, por último, la última etapa supone la redacción de la memoria.

En la siguiente tabla se recogen el número de horas dedicadas a cada una de las fases descritas anteriormente.

Etapa	Descripción de la tarea	Nº horas
Fase 1	Fase previa	220 horas
Fase 2	Algoritmos base	120 horas
Fase 3	Experimentos	492 horas
Fase 4	Memoria	120 horas

Tabla 10. Fases del proyecto

Partiendo de esto, deberemos considerar que en la realización del proyecto han participado dos personas, el ingeniero proyectante, con un salario de ingeniero junior, y el tutor, cuyo rango corresponde al de investigador. Así pues, la justificación de los costes de personal es la siguiente.

Personal	Horas de trabajo	€ / hora	Total (€)
Ingeniero Junior	850	11,5	9775
Investigador	102	20	2040
TOTAL		**11815 €**	

Tabla 11. Coste de personal

7.2.2. Costes de material

Por otra parte, debemos considerar los diferentes costes de material que conlleva la realización del proyecto. En primer lugar, los equipos informáticos utilizados para el desarrollo del sistema. En segundo lugar, la documentación empleada para tener conocimiento de estudios previos y de las técnicas existentes en el ámbito que nos ocupa. Y, por último, los gastos que supone la generación de la base de datos.

Concepto	Unidades	€ / unidad	Total (€)
Ordenador	1	500	500
Documentación	1	200	200
Resonancias magnéticas	172	150	25800
TOTAL		**26500 €**	

Tabla 12. Coste de material

7.2.3. Resumen de costes

Utilizando en conjunto todos los datos anteriores la siguiente tabla muestra el presupuesto total estimado para la ejecución del proyecto.

Concepto	Importe
Costes de personal	11815 €
Costes de material	26500 €
Base imponible	**38315 €**
IVA (21%)	8046,15 €
TOTAL	**46361,15 €**

Tabla 13. Presupuesto

Bibliografía

[1] Pujol, J., Soriano-Mas, C., Alonso, P., Cardoner, N., Menchón, J.M., Deus, J. Y Vallejo, J. (2004). Mapping structural brain alterations in obsessive-compulsive disorder. Arch Gen Psychiatry, 61(7): 720-730.

[2] Radua, J. y Mataix-Cols, D. (2009). Voxel-wise meta-analysis of grey matter changes in obsessive-compulsive disorder. The British J. of Psychiatry, 195(5): 393-402.

[3] Ford, J., Farid, H., Makedon, F., Flashman, L. A., McAllister, T. W., Megalooikonomou, V. y Saykin, A. J. (2003). Patient classification of MRI activation maps. En Proc. of the 6th Annual International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'03), páginas 58– 65.

[4] Shinkareva, S. V., Ombao, H. C., Sutton, B. P., Mohanty, A. y Miller, G. A. (2006). Classification of functional brain images with a spatio-temporal dissimilarity map. Neuroimage, 33(1):63–71.

[5] Costafreda, S. G., Chu, C., Ashburner, J. y Fu, C. H. (2009). Prognostic and diagnostic potential of the structural neuroanatomy of depression. PLoS ONE, 4:e6353.

[6] Ecker, C., Rocha---Rego, V., Johnston, P., Mourao Miranda, J. M., Marquand, A., Daly, E. M., Brammer, M. J., Murphy, C. y Murphy, D. G. (2010). Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. Neuroimage, 49(1):44–56.

[7] Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O. y The Alzheimer's Disease Neuroimaging Initiative (2011). Automatic classification of patients with alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage, 56(2):766–81.

[8] Soriano-Mas, C., Pujol, J., Alonso, P., Cardoner, N., Menchón, J. M., Harrison, B. J., Deus, J., Vallejo, J. y Gaser, C. (2007). Identifying patients with obsessive-compulsive disorder using whole-brain anatomy. Neuroimage, 35(3):1028–1037.

[9] Gómez-Verdejo, V., Verleysen, M. y Fleury, J. (2007). Information-theoretic feature selection for the classification of hysteresis curves. En Proc. 9th Intl. Work-Conference on Artificial Neural Networks, LNCS 4507, páginas 522– 529, San Sebastián, España.

[10] Gómez-Verdejo, V., Verleysen, M. y Fleury, J. (2009). Information-theoretic feature selection for functional data classification. Neuro-computing, 72:3580–3589.

[11] Gómez-Verdejo, V., Martínez-Ramón, M., Arenas- García, J., Gredilla, M. L. y Molina-Bulla, H. (2011). Support vector machines with constraints for sparsity in the primal parameters. IEEE Transactions on Neural Networks, 8:1269–1283.

[12] Castro, E., Martínez-Ramón, M., Pearlson, G., Sui, J., Calhoun, V.D., 'Characterization of groups using composite kernels and multi-source fMRI analysis data: Application to Schyzophrenia', Neuroimage, Vol. 58, No. 2, pp. 526-536, Sep, 2011.

[13] Parrado-Hernández, E., Gómez-Verdejo, V., Martínez-Ramón, M., Alonso, P., Pujol, J., Menchón, J., Cardoner, N. y Soriano-Mas, C. (2011). Identification of OCD-relevant brain areas through multivariate feature selection. En Proc. NIPS Workshop on Machine Learning and Interpretation in Neuroimaging, Sierra Nevada, España.

[14] Parrado-Hernández, E., Gómez-Verdejo, V., Martínez-Ramón, M., Shawe-Taylor, J., Alonso, P., Pujol, J., Menchón, J., Cardoner, N. y Soriano-Mas, C. (2012). Voxel selection in MRI through bagging and transduction: application to detection of obsessive compulsive disorder. En Proc. 2nd International Workshop on Pattern Recognition in Neuroimaging, PRNI 2012, London, UK.

[15] Cristianini, N. y Shawe-Taylor, J. (2000). An introduction to support vector machines. Cambridge University Press.

[16] Schlkopf, B. y Smola, A. J. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press.

[17] Guyon, I. y Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182.

[18] Guyon, I., Gunn, S., Nikravesh, M. y Zadeh, L. (editores) (2006). Feature Extraction, Foundations and Applications. Studies in Fuzziness and Soft Computing. Springer Verlag.

[19] Guyon, I., Weston, J., Barnhill S., Vapnik, V. Gene selection for cancer classification using support vector machines. - Machine learning, 2002 - Springer

[20] Zhu, J., Rosset, S., Hastie, T., Tibshirani, R. 1-norm Support Vector Machines. Neural Information Processing Systems, 2003. MIT Press.

[21] Soriano-Mas, C. Pre-procesado para VBM.

[22] Korbinian Brodmann. Website, 2010. http://es.wikipedia.org/wiki/Korbinian_Brodmann

[23] Tzourio-Mazoyer, N. Landeau, B. Papathanassiou, D. Crivello, F. Etar, O. Delcroix, N. Mazoyer, B. Joliot, M. Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. NeuroImage, 2002. 15: 273-289.

[24] Mitchell, T.M. Machine Learning. McGraw-Hill International Edition. 1997.

[25] Burges, C. A tutorial on support vector machines for pattern recognition. Bell Laboratories, Lucent Techonologies, USA. 1998. Data Mining and Knowledge Discovery, 2, 121-167.

[26] Valle Padilla, F. Implementación eficiente de clasificadores prior-SVM para MATLAB. Proyecto final de carrera. Universidad Carlos III de Madrid. Abril 2010.

[27] Blum, A. Langrey, P. Selection of relevant features and examples in machine learning. 1997. Artificial intelligence, 97(1-2): 245-271.